MODELING, ANALYSIS, AND OPTIMIZATION FOR WIRELESS NETWORKS IN THE PRESENCE OF HEAVY TAILS

A Thesis
Presented to
The Academic Faculty

by

Pu Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2013

MODELING, ANALYSIS, AND OPTIMIZATION FOR WIRELESS NETWORKS IN THE PRESENCE OF HEAVY TAILS

Approved by:

Professor Ian F. Akyildiz, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Raghupathy Sivakumar
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Chuanyi Ji
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Konstantinos Dovrolis
College of Computing
Georgia Institute of Technology

Professor Ye Li
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 6th August 2013

To my wife,

Xin Liu,

for her endless love and support.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Ian F. Akyildiz, for giving me such a precious opportunity to be his student and for his tremendous guidance, continuous support, and priceless encouragement through my PhD study. He not only trains me to conduct top-notch research, but also passes me his incredible life experience and great visions. He has been teaching me in many other ways, which makes me become much more persistent, determined, and ambitious, which I will benefit a lot from in my entire life.

I wish to express my sincere thanks to Dr. Chuanyi Ji, Dr. Ye Li, Dr. Raghupathy Sivakumar, and Dr. Konstantinos Dovrolis, who kindly agreed to serve in my dissertation defense committee. Their valuable comments and constructive suggestions have helped me to improve my dissertation in a solid way.

I want to thank all former and current members of the Broadband Wireless Networking (BWN) Laboratory for their support and friendship. It has been one of my most pleasant and memorabile life experience to be with the talented, caring, warmhearted, and passionate people in the BWN family.

Last but not the least, I would like to express my deep gratitude to my wife, Xin Liu, for her incredible support, encouragement, patience, and trust. Without her love and support, I could not to achieve what I have accomplished so far. .

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

Heavy-tailed distribution has been widely observed in a variety of computer and communication networks, which leads to the bursty nature of Internet and multimedia traffic, the highly variable channel condition, and the irregular mobility pattern of network users. Different from the conventional light-tailed one, heavy-tailed network environment exhibits extremely high variability and dynamics that can fundamentally change the way in which wireless networks are conceived, designed, and operated.

This thesis is concerned with modeling, analysis, and optimization of wireless networks in the presence of heavy tails. First, a novel traffic model is proposed, which captures the inherent relationship between the traffic dynamics and the joint effects of the mobility variability of network users and the spatial correlation in their observed physical phenomenon. Then, the statistical attributes of the proposed model are analyzed, which establish the conditions under which user mobility associated with spatial correlation can lead to heavy-tailed traffic. To mitigate the bursty nature of heavy-tailed traffic, a mobility-aware traffic shaping scheme is proposed to actively regulate the network traffic by coordinating users' mobility patterns.

Under heavy-tailed environment, the fundamental network performance limits in terms of delay, stability, and connectivity are first analyzed. Then, the optimal algorithms to approach these performance limits are proposed. More specifically, the asymptotic delay distribution of wireless users is analyzed under different traffic patterns and spectrum conditions, which reveals the critical conditions under which wireless users can experience heavy-tailed delay with significantly degraded QoS performance. To encounter this problem, multi-channel multi-radio solutions, including

spectrum mobility and multi-radio diversity, are developed. The proposed solutions aim to yield the optimal QoS performance by exploiting the temporal diversity of wireless channels in such a way that the heavy-tailed delay is mitigated to the maximum extent.

Based on the delay analysis, the fundamental impact of heavy-tailed environment on network stability is studied. Specifically, a new network stability criterion, namely moment stability, is introduced to better characterize the stability performance in the presence of heavy-tailed traffic. Then, an asymptotic queueing analysis is performed to reveal the critical conditions under which there exists a feasible scheduling policy to achieve moment stability. Utilizing this analysis, a maximum-weight-β scheduling algorithm is proposed, which associates each queue with a different parameter β and makes the scheduling decision based on the queue lengths raised to the β-th power. It is proven that the maximum-weight-β scheduling algorithm is throughput-optimal in the sense that it stabilizes the network for any arrival rates in the stability region.

Furthermore, the impact of heavy-tailed spectrum on network connectivity is investigated. Towards this, a new connectivity criterion, namely delay-bounded connectivity, is introduced, which simultaneously ensures the existence of routing paths and the finiteness of the average delay and jitter along these paths. Then, the necessary conditions on the existence of delay-bounded connectivity are derived. To enhance network connectivity degraded by the spectrum burstiness, mobility-assisted data forwarding schemes are exploited, which dramatically increase the information delivery opportunities by exploiting the spatial diversity of the wireless spectrum and the opportunistic contacts among mobile devices. Accordingly, as an important design parameter for all mobility-assisted data forwarding schemes, the critical mobility radius is derived, which is a critical threshold on the maximum radius the network user can reach, above which delay-bounded connectivity is guaranteed. Moreover, the end-to-end latency of the mobility-assisted data forward schemes is analyzed, which

exhibits asymptotic linearity in the initial distance between mobile users.

CHAPTER I

INTRODUCTION

1.1 Background

The emerging Internet and multimedia applications, such as voice over Internet Protocol (VoIP), telemedicine, online gaming, video conferencing, and multimedia surveillance, are expected to become dominant in current and next-generation wireless networks. Providing substantial levels of network quality of service (QoS) for these applications in wireless domain is challenging because of the bursty nature of Internet and multimedia traffic, the highly variable channel condition, and the mobility of wireless network users. Specifically, significant empirical evidence establishes that the burst duration of multimedia and Internet traffic, the channel occupancy time of wireless links, and the travel distance of mobile users can follow heavy-tailed distribution [29] [44][63][32]. This distribution, compared with the conventional light-tailed one, e.g., exponential distribution and Poisson distribution, exhibits significantly different statistical attributes that can completely change the way how wireless networks are conceived, designed, and operated. The objective of this thesis is to develop effective traffic models, analyze fundamental performance limits, and design optimal control schemes for wireless networks in the presence of heavy tails.

The performance of wireless networks heavily depends on the statistical properties of the traffic arrival process. The seminal work of Leland et al. [29] established that Internet traffic is heavy tailed and thus exhibits a property of correlation over a wide range of time scales. The similar heavy tail properties are also widely observed in multimedia traffic such as variable bit rate (VBR) video streams [41]. Since heavy tail traffic behaves extremely bursty on a wide range of time scales, the impact of heavy

tail traffic on network performance is significantly different from that of light tailed traffic, e.g., Poisson traffic. Particularly, heavy tail traffic can induce much larger delay and higher packet loss rate than light tail traffic. These important consequences motivate a great deal of research efforts on studying the origins of heavy tail traffic. For example, the statistical analysis on real network traffic traces verifies that the tail distribution of the file size on Internet web servers is heavy tailed and thus decays much slower than exponentially [29]. This fact gives rise to very long traffic bursts with non-negligible probability and thus causes long range dependent behavior of Internet traffic.

In wireless networks, the traffic profile of the users is also considerably affected by user mobility. Specifically, the traffic initiated by a mobile user could be associated with different locations the user has visited. The traffic content generated at these locations can exhibit a certain level of correlation, which may excessively depend on the mutual distance between these locations [51]. Surprisingly, recent studies confirm that the travel distance of mobile users follows heavy-tailed distribution [44], in sharp contrast to the conventional light-tailed assumptions. Therefore, it is important to assess whether and how such mobility pattern along with the spatial correlation impacts the statistical properties of network traffic, in particular the rise of heavy tailed properties. Accordingly, the revealed impact can be exploited to develop effective traffic models and traffic engineering solutions.

Besides the heavy-tailed nature of network traffic, the dynamically changing wireless channel is another challenge of providing desired QoS performance for wireless and mobile applications. The frequently disconnected wireless links not only cause severe packet loss but also invoke undesirable retransmissions, which can further induce unbounded delay and jitter even if the transmitted messages are of bounded size [47]. What is more important, recent empirical results show that the channel occupancy time of both WiFi and cellular networks exhibits heavy-tailed statistics,

which inevitably lead to extremely high channel variations [63][32]. Such channel behavior has been demonstrated to profoundly impact the performance of spectrum sensing operations, which serve as the building block for the network control functions such as interference management and medium access control [63]. Despite its importance, the impact of heavy-tailed traffic and channel models on the fundamental performance limits of wireless networks is still under-exploited. To fill this gap, the critical bounds of the key network attributes such as delay, stability, and connectivity need to be analyzed. By this analysis, wireless networks can be better designed to meet the increasing demands for the emerging multimedia and Internet applications in wireless domain.

Achieving the performance limits of wireless networks requires the optimal design of network control functions at nearly all layers of the protocol stack such as multi-channel diversity and power control at physical layer, medium access control or scheduling at link layer, routing decisions at network layer, and congestion and admission control at transport layer. Specifically, to minimize transmission delay, multi-radio multi-channel diversity has been widely employed in the majority of wireless networks by exploiting the transmission opportunities at multiple channels [27]. In addition, to maintain desired network stability while maximizing the utilization of limited spectrum resources, throughput-optimal scheduling policies have been proposed with an objective to support the largest set of traffic rates that is allowed by a given network [38][48]. Moreover, to improve network connectivity, the opportunistic contacts among mobile devices have been exploited to increase the information dissemination opportunities in the wireless networks with frequently impaired links [22]. However, all these network control solutions were initially designed under the light-tailed assumptions and their effectiveness in the presence of heavy tails is questionable. This necessitates the design of new optimal network control solutions that fundamentally depart from the conventional ones.

1.2 Research Objectives and Solutions

Heavy-tailed traffic has been widely observed in a variety of computer and communication networks. Different from the conventional light-tailed traffic, heavy-tailed traffic exhibits high burstiness or strong dependence over a long range of time scales. Such highly bursty nature can fundamentally change the way in which wireless networks are conceived, designed, and operated.

This proposal is concerned with modeling, analysis, and optimization of wireless networks in the presence of heavy-tailed traffic. First, we propose a novel traffic model, which captures the inherent relationship between the heavy-tailed traffic and the joint effects of the mobility variability of network users and the spatial correlation in their observed physical phenomenon. Then, under heavy-tailed environment, we analyze the fundamental network performance limits in terms of latency, stability, and connectivity. Then, we propose optimal algorithms spanning different protocol layers to approach these performance limits. At physical layer, we provide an asymptotic analysis of the transmission delay under dynamically changing spectrum with heavy-tailed statistics. This analysis reveals that the promising dynamic spectrum access schemes inevitably induce heavy-tailed delay that leads to significantly degraded QoS. To encounter this problem, we propose optimal multi-radio multi-channel solutions to mitigate heavy-tailed delay. Based on the delay analysis, at MAC layer, we study the impact of heavy-tailed traffic on the stability performance of wireless networks and accordingly design throughput-optimal scheduling policies with an objective to maximize network throughput, while maintaining network stability in the presence of heavy-tailed traffic. At network layer, we give a percolation-based connectivity analysis under bursty spectrum with heavy-tailed behavior. To enhance network connectivity impaired by the spectrum burstiness, we exploit mobility-assisted data forwarding schemes, which dramatically increase the information delivery paths by utilizing the spatial diversity of the wireless spectrum and the opportunistic contacts

among mobile devices.

1.2.1 On the New Origins of Heavy-tailed Traffic

The design of effective traffic models plays a key role in evaluating network performance and developing efficient network control schemes. In a wireless network, a mobile node may inject to the network the similar traffic at the proximal locations because their collected information at these location could be spatially correlated. This yields temporal correlation in the network traffic. If this correlation decays slowly as the time lag increases, then the traffic could exhibit long range dependent (LRD) or heavy-tailed behavior. The traffic is LRD if its autocorrelation function follows a power-law form as the lag approaches infinity. Previous works [29][16] have shown that LRD traffic leads to fundamentally different impact on network performance and protocol design, compared with the traditional traffic. To design effective and efficient protocols for wireless applications, it is essential to investigate the relationship between the LRD behavior and the dynamics induced by mobility and spatial correlation.

Although a large number of traffic models have been proposed, limited research efforts have been devoted to the traffic modeling for the emerging wireless sensor and actor networks (WSANs). The long-lasting, low-cost, and cooperative nature of WSANs facilitates a wide range of civilian and military applications such as environmental monitoring, industrial control, e-health system, battle field surveillance, and automatic border patrol. In WSANs, there generally exist two types of nodes: static sensor nodes, which are deployed in a large area of interests, and mobile actors, which roam around in this area. The mobile actors can retrieve the sensing data from different static sensor nodes, exchange the data with other actors, and initiate certain actions accordingly. Although some traffic models have been proposed for the conventional wireless sensor networks only consisting of static sensor nodes [13][35],

none of them can characterize the traffic features of the mobile actors in WSANs. This fact necessitates the design of new structural traffic models that explicitly capture the correlation between the unique mobility feature of the actors and their traffic patterns.

In this work, we reveal the new origins of heavy-tailed traffic by studying the relationship between the spatial correlation and the temporal dependency [55][52]. Towards this, we propose a novel traffic modeling scheme that captures the statistical patterns of spatial correlation and mobility. Then, the statistical attributes of the proposed model are analyzed, which establish the conditions under which user mobility associated with spatial correlation can lead to heavy-tailed traffic. To mitigate the bursty nature of heavy-tailed traffic, we propose new traffic shaping protocols based on mobility coordination to actively shape the traffic so that the resulting traffic can follow the desired characteristics.

1.2.2 Asymptotic Delay Analysis under Heavy-tailed Environment

Transmission delay, as one of the key QoS metrics, has been widely studied for classical communication network paradigms. In this work, we provides an asymptotic analysis of the transmission delay under dynamic spectrum access schemes. Dynamic spectrum access (DSA) is an emerging technique that allows the secondary users (SUs) to share the spectrum in an opportunistic manner [2]. Using such scheme, the SUs can access the unoccupied spectrum during idle periods of the primary users (PUs), and stop transmissions when the PU channels become busy. The achievable Quality of Service (QoS) performance of secondary users is significantly affected by the dynamically changing PU traffic and the resource allocation policies used by the SUs.

We first investigate the delay performance when only a single PU channel is utilized. Specifically, it is shown that DSA induces only light-tailed delay as long as both the busy time of PU channels and the message size of SUs are light tailed. On

the contrary, if either the busy time or the message size is heavy tailed, then the SUs' transmission delay is heavy tailed. For this case, we prove that if one of the busy time or the message size is light tailed and the other is regularly varying with index α, then the transmission delay is regularly varying with the same index α. As a consequence, the delay has an infinite variance provided $\alpha < 2$ and an infinite mean provided $\alpha < 1$. This implies that even if transmitting messages with finite mean size, SUs can experience extremely high delay variation and even stochastically zero throughout. Furthermore, if both the busy time and the message size are regularly varying with index α and β, respectively, then the tail distribution of the delay is as heavy as the tail distribution of either the message size or the busy time, whichever has the smaller index.

Moreover, we investigate the benefits of exploiting the transmission opportunities on multiple PU channels [57][54]. More specifically, we consider two multiple-channel access schemes, namely, spectrum mobility and multi-radio diversity. Under spectrum mobility, if a PU appears in a channel currently used by an SU, the SU vacates the channel immediately and continues its transmission in another idle channel [2]. Under multi-radio diversity, an SU is equipped with multiple radio interfaces so that it can simultaneously access multiple channels. We show that compared with the case in which only a single channel is used, both spectrum mobility and multi-radio diversity can mitigate the degree of heavy-tailed delay by increasing the orders of its finite moments.

1.2.3 Throughout-optimal Scheduling Algorithms under Heavy-tailed Traffic

While transmission delay is a single-user performance metric, stability is a network-wide one. Generally speaking, a network is stable if there exists a feasible scheduling algorithm under which each network user has finite time-average expected queue length [38][48]. Accordingly, the network stability region is defined as the closure

of the set of all arrival rate vectors for which the queues of all wireless users can be stabilized by a feasible scheduling policy. Moreover, a scheduling policy is throughput optimal if it stabilizes the system for any arrival rates in the stability region. Although network stability regions and throughput-optimal scheduling algorithms have been investigated in different network scenarios, they are generally derived under light-tailed channel and traffic models [38][48]. However, recent research on queueing theory [33] shows that when two queues share the same server under the maximum weight scheduling algorithm, which is proven to be throughput optimal under light-tailed traffic arrivals, the queue with light-tailed traffic can experience unbounded average delay if the other queue has heavy-tailed traffic with unbounded delay variance.

The above observations motivate us to investigate network stability from a new perspective [58][56]. Specifically, we introduce a new stability criterion, namely *moment stability*, which requires that all the network users with light-tailed arrival traffic always have bounded queueing delay with finite mean and variance. Although moment stability is a desirable property to promise QoS guaranteed applications, the conventional scheduling policies, which are effective under the light-tailed traffic, have difficulty in achieving moment stability in the presence of heavy tails. Consequently, new scheduling policies, which cope with the heavy-tailed network environment, have to be developed. Towards this end, we first study the queue length asymptotics under two basic channel access solutions: the exclusive access policy and the shared access policy. The former policy allows a SU has exclusive access to the PU channel without competing with other SUs, while the latter policy requires all SUs to share the PU channel. Based on this study, we reveals the necessary conditions under which there exists a feasible scheduling policy to achieve moment stability. Accordingly, we propose the maximum-weight-β scheduling algorithm, which associates each queue with a different parameter β and makes the scheduling decision based on the queue lengths raised to the β-th power. In particular, we show that maximum-weight-β

scheduling algorithm is throughput-optimal in the sense that it can maximize the network throughput, while maintaining moment stability. More specifically, we prove its throughput optimality by giving an asymptotic queueing analysis, which shows that there always exists a feasible set of β parameters such that the maximum weight-β scheduling yields the best asymptotic queueing performance by letting each queue have the lightest possible tail and consequently the highest possible order of finite moments.

1.2.4 Mobility Improves Delay-bounded Latency with Heavy-tailed Spectrum

Similar to network stability, connectivity is also a network-wide attribute that has to be maintained for reliable communications between transmitting and receiving parties in a network. Conventionally, there exist two types of connectivity: full connectivity and percolation-based connectivity. Specifically, full connectivity ensures that each pair of nodes in the network is connected by at least one path. However, for wireless networks, this connectivity criterion is overly restrictive or difficult to achieve because of the complicated radio environment, unplanned network topology, and severe impacts from coexisting networks. Different from full connectivity, percolation-based connectivity only requires a network to contain an extremely large connected component such that each node in this component can connect with an extremely large number of other nodes [15] [25][53].

Although percolation based connectivity can characterize the existence of routing paths between network devices, it does not indicate the end-to-end quality of service, such as delay and jitter. We consider a heterogeneous network setting, where there exist two networks: the primary network and the secondary network, where the primary network has the priority to access the spectrum. As implied by our delay and queueing analysis, under heavy-tailed spectrum activities, as the density of primary users increases, the primary network can generate extremely large interference

region within which the secondary users will experience significantly degraded delay performance. Therefore, a latency-oriented connectivity definition is more meaningful. Towards this, we first define a new connectivity criterion, namely delay-bounded connectivity, which simultaneously ensures the existence of routing paths and the finiteness of the average delay and jitter along these paths.

In this work, we study the fundamental impact of heavy-tailed spectrum activities on the delay-bounded connectivity as well as how and to what extent mobility can mitigate such impact [60][59][53]. More specifically, we show that such heavy tailed spectrum activities significantly degrade the connectivity of secondary network. Specifically, it is proven that if the busy time of primary users is heavy tail distributed, there always exists a critical density λ_p such that if the density of primary users is larger than λ_p, the secondary network cannot achieve delay-bounded connectivity surely. To encounter this problem, the mobility of secondary users is utilized to exploit the spatial diversity of the spectrum availability through the opportunistic contacts of mobile users. In particular, we prove that there exists a critical threshold on the maximum radius the secondary user can reach, above which the secondary network can already achieve delay-bounded connectivity, independent of primary network impact such as node density and activities. Moreover, we study the latency performance of the mobility-assisted data forward schemes, which shows that their yielded end-to-end latency scales linearly in the initial distance between two mobile users.

1.3 Organization of the Thesis

This thesis is organized as follows.

In Chapter 2, we first introduce the mathematical background of heavy tail distribution and long range dependent traffic. Then, we present and prove some useful lemmas that are applied throughout the thesis.

In Chapter 3, we describe our proposed traffic model in details. Then, we analyze the statistical attributes of the proposed model and establish the conditions under which user mobility associated with spatial correlation can lead to heavy-tailed traffic. Then, we propose a mobility-aware traffic shaping scheme that can effectively mitigate the bursty nature of heavy-tailed traffic.

In Chapter 4, we first give an asymptotic delay analysis of wireless users under different traffic patterns and spectrum conditions. Based on this analysis, we reveal the critical conditions under which wireless users can experience heavy-tailed delay with significantly degraded QoS performance. Then, to mitigate the heavy-tailed delay. we propose two multi-channel multi-radio solutions, including spectrum mobility and multi-radio diversity, and analyze their delay performance, respectively.

In Chapter 5, we first formally define a new network stability criterion, namely moment stability. Then, we give an asymptotic queueing analysis to reveal the critical conditions on the existence of moment stability in presence of heavy tails. Based on the analysis, we propose the maximum-weight-β scheduling algorithm and prove its throughput optimality.

In Chapter 6, we first define a new connectivity criterion, namely delay-bounded connectivity. Then, we reveal the sufficient conditions under which delay-bounded connectivity is not achievable due to the heavy-tailed nature in the radio spectrum. Next, we introduce mobility-assisted data forwarding schemes and derive the critical mobility radius above which delay-bounded connectivity is guaranteed surely. Moreover, we analyze the end-to-end latency of the mobility-assisted data forward schemes.

In Chapter 7, we draw the main conclusions and summarize the future work.

CHAPTER II

HEAVY TAIL DISTRIBUTION AND LONG RANGE DEPENDENCE

In this chapter, we first introduce the mathematical background of heavy tail distribution and long range dependent traffic. Then, we present and prove some useful lemmas that are applied throughout the thesis.

2.1 Heavy Tailed Distribution

In this thesis we use the following notations. For any two real functions $a(t)$ and $b(t)$, we let $a(t) \sim b(t)$ denote $\lim_{t\to\infty} a(t)/b(t) = 1$. We say that $a(t) \lesssim b(t)$ if $\limsup_{t\to\infty} a(t)/b(t) \leq 1$, and $a(t) \gtrsim b(t)$ if $\liminf_{t\to\infty} a(t)/b(t) \geq 1$. Furthermore, we say that $a(t) = o(b(t))$ if $\lim_{t\to\infty} a(t)/b(t) = 0$. In addition, for any two non-negative r.v.s X and Y, we say that $X \leq_{a.s.} Y$ if $X \leq Y$ almost surely, and $X \leq_{s.t.} Y$ if X is stochastically dominated by Y, i.e., $P(X > t) \leq P(Y > t)$ for all $t \geq 0$. We say $X \stackrel{d}{=} Y$ if X and Y are equal in distribution. Also, let $F(x) = P(X \leq x)$ denote the cumulative distribution function (cdf) of a non-negative r.v. X. Let $\overline{F}(x) = P(X > x)$ denote its tail distribution function.

Definition 1. *A r.v. X is heavy tailed (HT) if for all $\theta > 0$*

$$\lim_{x\to\infty} e^{\theta x}\overline{F}(x) = \infty, \tag{1}$$

or, equivalently, if for all $z > 0$

$$E[e^{zX}] = \infty. \tag{2}$$

Definition 2. *A r.v. X is light tailed (LT) if there exists $\theta > 0$ such that*

$$\lim_{x\to\infty} e^{\theta x}\overline{F}(x) = 0, \tag{3}$$

or, equivalently, if there exists $z > 0$ such that

$$E[e^{zX}] < \infty. \tag{4}$$

Remark 1. *Generally speaking, a r.v. is HT if its tail distribution decreases slower than exponentially. Some typical HT distributions include Pareto, log-normal, Bur, and Weibull (with shape parameter less than 1) distributions. On the contrary, a r.v. is LT if its tail distribution decreases exponentially or faster. Some typical LT distributions cover exponential, Gamma, and Weibull (with shape parameter larger than 1) distributions. A key characteristic that distinguishes HT r.v.s from LT ones is that the moment generating function of any HT r.v. X is infinite, i.e., $E(e^{zX}) = \infty, \forall\ z > 0$.*

Based on the existence of the moments, we define the tail index of a non-negative random variable.

Definition 3. *The tail index $\kappa(X)$ of a nonnegative random variable X is defined by*

$$\kappa(X) = \sup\{k \geq 0 : E[X^k] < \infty\}. \tag{5}$$

Remark 2. *The tail index specifies the threshold order above which a random variable has infinite moments. Some HT distributions, such as Pareto, have finite tail index, which leads to infinite moments of certain orders, Some HT distribution, such as log-normal, have infinite tail index and therefore possesses finite moments of all orders. In this work, we focus on heavy tail distributed random variables with finite tail index because they can effectively characterize lots of network attributes such as the frame length of variable bit rate (VBR) traffic, the session duration of licensed users in WLANs, and files sizes on internet severs [32] [41].*

The following Lemma presents the sufficient condition regarding the existence of finite tail index for a r.v. X [11].

Lemma 1. *A nonnegative r.v. X has $\kappa(X)$ if and only if the tail distribution of X satisfies*

$$\lim_{t\to\infty} \frac{\log[P(X > t)]}{\log t} = -\kappa(X). \tag{6}$$

An important subclass of HT distributions is the class of regularly varying distributions [4]. Its definition involves the slowly varying function which is defined as follows.

Definition 4. *A measurable positive function $\mathcal{L}(x)$ defined in some interval $[a, \infty)$ is called slowly varying if for all $y > 0$*

$$\lim_{x\to\infty} \frac{\mathcal{L}(yx)}{\mathcal{L}(x)} = 1 \tag{7}$$

For example, a constant and a logarithmic function are both slowly varying functions.

Lemma 2. *(Properties of slowly varying function [4])*

1. *If $\mathcal{L}(x)$ varies slowly, $\lim_{x\to 0} \log(\mathcal{L}(x))/\log x = 0$.*
2. *If $\mathcal{L}(x)$ varies slowly, so does $(\mathcal{L}(x))^a$ for every $a \in \mathbb{R}$.*
3. *If $\mathcal{L}_1(x)$ and $\mathcal{L}_2(x)$ vary slowly, so do $\mathcal{L}_1(x) + \mathcal{L}_2(x)$ and $\mathcal{L}_1(x)\mathcal{L}_2(x)$.*

Definition 5. *A r.v. X is called regularly varying with index $\alpha > 0$, denoted by $X \in \mathcal{RV}(\alpha)$, if*

$$\overline{F}(x) \sim x^{-\alpha}\mathcal{L}(x), \tag{8}$$

where $L(x)$ is a slowly varying function.

Remark 3. *Regularly varying distributions are a generalization of power law distributions. The index α indicates how heavy the tail distribution is, where smaller values of α imply heavier tail. Moreover, for a r.v. $X \in \mathcal{RV}(\alpha)$, the exact values of α determine whether the moments of X are bounded or not. This is explained in the following lemma.*

Lemma 3. *For any r.v. $X \in \mathcal{RV}(\alpha)$, the moments of order $m > \alpha$ is unbounded, i.e.,*

$$E[X^m] = \infty, \quad \forall m > \alpha. \tag{9}$$

Remark 4. *In particular, for any r.v. $X \in \mathcal{RV}(\alpha)$, if $\alpha < 1$, X has an infinite mean. If $1 < \alpha < 2$, X has a finite mean but an infinite variance.*

The following preliminary Lemmas regarding regularly varying and light tailed distributions are also useful in this thesis. We first state the Lemmas, followed by their proofs.

Lemma 4. *Let $X \in \mathcal{RV}(\alpha)$ and $Y \in \mathcal{RV}(\beta)$. If $\alpha > \beta$, then*

$$P(X > at) = o(P(Y > bt))$$

with $a > 0$ and $b > 0$.

Lemma 5. *Let X and Y be non-negative random variables. If $X \in \mathcal{RV}(\alpha)$ and $P(Y > t) = P(X > bt)$ with $b > 0$, then $Y \in \mathcal{RV}(\alpha)$*

Lemma 6. *Let X be LT and $Y \in \mathcal{RV}(\alpha)$. Then,*

$$P(X > at) = o(P(Y > bt))$$

with $a > 0$ and $b > 0$.

Lemma 7. *Let X and Y be non-negative random variables. If $P(Y > t) = P(X > a(t + b))$ with $0 < a < \infty$ and $0 < b < \infty$, then Y is LT provided X is LT.*

Let $\{Y_i\}_{i\geq 1}$ be non-negative i.i.d. random variables independent of the non-negative random variable N. Define $S_N := \sum_{i=1}^{N} Y_i$. We have following Lemma 8 [17] and 9.

Lemma 8. *1. Assume $Y_1 \in \mathcal{RV}(\alpha)$, $E[N] < \infty$ and $P(N > t) = o(P(Y_1 > t))$. Then,*

$$P(S_N > t) \sim E[N]P(Y_1 > t).$$

2. *Assume* $N \in \mathcal{RV}(\alpha)$, $E[Y_1] < \infty$, *and* $P(Y_1 > t) = o(P(N > t))$. *Moreover, assume that* $E[N] < \infty$ *if* $\alpha = 1$. *Then,*

$$P(S_N > t) \sim P(N > (E[Y_1])^{-1}t).$$

Lemma 9. *Assume* $N, Y_1 \in \mathcal{RV}(\alpha)$ *with* $E[N] < \infty$. *Let* $P(N > t) = t^{-\alpha}\mathcal{L}_1(t)$ *and* $P(Y_1 > t) = t^{-\alpha}\mathcal{L}_2(t)$. *Then,*

$$P(S_N > t) \sim E[N]P(Y_1 > t) + (E[Y_1])^{\alpha}P(N > t). \tag{10}$$

Lemma 10. *Properties of LT Distributions [37]*

1. *If* X *and* Y *are non-negative LT random variables, then* $X + Y$ *is LT.*

2. *Let* $\{X_i\}_{i\geq 1}$ *be i.i.d. LT random variables, and* N *be integer LT random variable. Then, the random sum* $\sum_{i=1}^{N} X_i$ *is LT.*

3. *Let* L *be a non-negative random variable and* $\{X_i\}_{i\geq 1}$ *be non-negative i.i.d. random variables independent of* L *and satisfying* $P(X_i > 0) > 0$. *If* L *is LT, so is* $\inf\{n : \sum_{i=1}^{n} X_i \geq L\}$.

2.2 Long Range Dependent Traffic

The traffic is called long range dependent if its autocorrelation function follows a power-law form as the time lag approaches infinity

$$\rho(\tau) \to c_p\tau^{\beta-1}, as\tau \to \infty$$

where c_p is a positive constant and $0 < \beta < 1$ is the fractal exponent. The quantity $H = (\beta + 1)/2$ is referred to as the Hurst parameter, which characterizes the speed of decay of the autocorrelation function. Note that LRD traffic has slowly decaying autocorrelation function with $0.5 < H < 1$, since $0 < \beta < 1$. By contrast, the short-range dependent traffic, e.g., Poisson traffic, has the exponentially decaying autocorrelation function.

2.3 Proofs of the Preliminary Lemmas

Proof of Lemma 4 to 7. The proof follows easily from the definitions of LT and HT r.v.s.. □

Proof of Lemma 9. We use techniques similar to those used in [17] to prove that the lower and upper bounds in (10) asymptotically coincide. For every fixed n_0 we obtain

$$\begin{aligned} P(S_N > t) = &\sum_{n=1}^{n_0} P(N=n)P(S_n > t) \\ &+ \sum_{n=n_0}^{\infty} P(N=n)P(S_n > t) \end{aligned}$$

Since $Y_1 \in \mathcal{RV}(\alpha)$, Y_1 is subexponentially distributed. By the subexponentiality of Y_1 and the independence of N and Y_1, we obtain

$$\begin{aligned} \sum_{n=1}^{n_0} P(N=n)P(S_n > t) &\sim \sum_{n=1}^{n_0} P(N=n)nP(Y_1 > t) \\ &\sim E[N]P(Y_1 > t), n_0 \to \infty \end{aligned}$$

For any $1 > \delta > 0$, we obtain for large enough t

$$\begin{aligned} &\sum_{n=n_0+1}^{\infty} P(N=n)P(S_n > t) \\ =&\left(\sum_{n=n_0+1}^{t(1-\delta)/E[Y_1]} + \sum_{n=t(1-\delta)/E[Y_1]}^{\infty} \right) P(N=n)P(S_n > t) \\ :=&I + II. \end{aligned}$$

For term II, we obtain

$$\begin{aligned} II = &\left(\sum_{n=t(1-\delta)/E[Y_1]}^{t(1+\delta)/E[Y_1]} + \sum_{n=t(1+\delta)/E[Y_1]}^{\infty} \right) P(N=n)P(S_n > t) \\ :=&J_1 + J_2 \end{aligned} \tag{11}$$

By the law of large numbers and letting $\delta \downarrow 0$, we obtain

$$\begin{aligned} J_1 \leq & \sum_{n=t(1-\delta)/E[Y_1]}^{t(1+\delta)/E[Y_1]} \left(P(N=n) P(\sum_{i=1}^{t(1+\delta)/E[Y_1]} Y_i > t) \right) \\ \sim & P\left(N > \frac{t(1-\delta)}{E[Y_1]} \right) - P\left(N > \frac{t(1+\delta)}{E[Y_1]} \right) = o(1) \end{aligned}$$

For J_2, we have

$$J_2 \leq \sum_{n=t(1+\delta)/E[Y_1]}^{\infty} P(N=n) \sim P\left(N > \frac{1+\delta}{E[Y_1]} t \right) \tag{12}$$

and by the law of large numbers,

$$\begin{aligned} J_2 \geq & \sum_{n=t(1+\delta)/E[Y_1]}^{\infty} P(N=n) P\left(\sum_{i=1}^{t(1+\delta)/E[Y_1]} Y_i > t \right) \\ \sim & P\left(N > \frac{1+\delta}{E[Y_1]} t \right) \end{aligned} \tag{13}$$

. Combining (12) and (13) and letting $\delta \downarrow 0$, we have

$$J_2 \sim P\left(N > \frac{t}{E[Y_1]} \right) \sim (E[Y_1])^{\alpha} P(N > t) \tag{14}$$

For term I, we have

$$I = \sum_{n=n_0+1}^{t(1-\delta)/E[Y_1]} P(N=n) P(S_n - nE[Y_1] > t - nE[Y_1])$$

Since $n < t(1-\delta)/E[Y_1]$, we obtain that $y := t - nE[Y_1] > nE[Y_1]((1-\delta)^{-1} - 1)$. By large deviations theory [6] [36], it follows that for any $\varepsilon > 0$

$$\lim_{n \to \infty} \sup_{y > \varepsilon n} \left| \frac{P(S_n - nE[Y_1] > y)}{nP(Y_1 > y)} - 1 \right| = 0$$

which implies that there exists some positive constant C such that

$$\lim_{n_0 \to \infty} \limsup_{t \to \infty} I \leq \lim_{n_0 \to \infty} C \sum_{n=n_0+1}^{\infty} P(N=n) n P(Y_1 > y) = 0$$

This, in conjunction with (11), (12) and (14), completes the proof. □

CHAPTER III

ON THE NEW ORIGINS OF HEAVY-TAILED TRAFFIC

3.1 Introduction

In the last few years, a significant number of research efforts have been devoted to the study of developing wide area distributed wireless sensor networks (WSNs) with self-organizing capabilities to cope with sensor node failures, changing environmental conditions, and sensing application diversity [3]. In particular, wireless sensor and actor network (WSAN) emerges as a promising candidate to support self-organizing mechanisms, enhancing adaptability, scalability, and reliability [5][42][19].

In a WSAN, there generally exist two types of nodes: the static sensor nodes, which are deployed in a large area of interests; and mobile agents, which roam around within this area. The mobile agents can retrieve the sensing data from the static sensor nodes, and exchange the data with other mobile agents or transmit the data directly to the remote sink. Such hybrid network scenario, which is illustrated in Figure 1, facilitates a wide range of civilian and military applications such as battlefield surveillance, nuclear, biological and chemical attack detection, and environmental monitoring [1]. However, the traffic model in WSAN has not been investigated yet. Conventionally, Markovian or Constant Bit Rate (CBR) traffic model is generally assumed in WSNs without any discussion as to whether this is appropriate or not. However, since WSNs are an application specified networking paradigm, different applications or network scenarios can yield different traffic patterns. For example, the WSNs, which are used for intrusion detection and medical applications, can generate traffic significantly different from the conventionally used Poisson or CBR traffic model [12][12]. Therefore, in this chapter, we aim to investigate the influence of two

key factors in WSANs, i.e., node mobility and spatial correlation of sensing observations, on the traffic patterns of the mobile nodes. More specifically, we will study whether the joint effects of the two factors induce Long Range Dependent (LRD) traffic because such traffic can lead to fundamentally different impact on network performance and protocol design, compared with the traditional Markovian traffic [29][40].

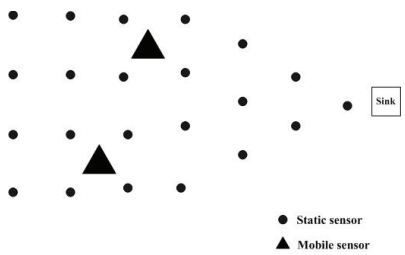

Figure 1: Wireless sensor and actor network.

The seminal work of Leland et al. in 1994 [29] established that Ethernet traffic exhibits a property of correlation over many different time scales and suggested that simple LRD models could be applied to effectively capture these correlations. Since then, a great number of research efforts have been devoted to the study of LRD behavior because of the impact of LRD on network performance and resource allocation, which exhibits characteristics significantly different from Markovian traffic. For example, LRD traffic can induce much larger delays than predicted by traditional queuing models. Furthermore, buffering, as a resource allocation strategy, becomes ineffective with LRD input traffic in the sense of incurring a disproportional penalty in queuing delay compared with the gain in reduced packet loss probability.

The important consequences of LRD necessitate us to answer the following questions regarding (1) whether the joint effects of mobility and spatial correlation lead to LRD traffic, and (2) if the answer is yes, how they affect the Hurst parameter which is used to measure the intensity of LRD in the traffic. To answer these questions, we construct an analytical traffic model that incorporates the statistical patterns of

node mobility and spatial correlation, and study the impact of these attributes on traffic statistics, which leads to several novel findings [55] [52]. These findings provide valuable new insights into questions related to the design of efficient and effective protocols for WSANs. The contributions made in this chapter conclude:

1. An analytical traffic model is proposed whose parameters are related to the main attributes of WSANs (e.g., mobility and spatial correlation). This model advances an explicit explanation of the impact of these attributes on the statistical patterns of the network traffic.

2. We find that the joint effects of mobility and spatial correlation can lead to the pseudo-LRD traffic, whose autocorrelation function approximates that of the LRD traffic with the Hurst parameter up to a certain cutoff time lag.

3. We show that the Hurst parameter is completely determined by the mobility variability and the degree of spatial correlation. Particularly, higher mobility variability and smaller spatial correlation could give rise to the burstier traffic with larger Hurst parameter. Conversely, lower variance in mobility and larger spatial correlation can lead to non-bursty traffic.

4. We demonstrate that the traffic at the relay node has the same Hurst parameter as the single node, but exhibit much more fluctuations per unit time than the single node traffic.

5. We propose a novel traffic shaping protocol, which can effectively reduce the traffic burstiness by properly coordinating the movement of the sensor nodes.

The rest of this chapter is organized as follows. In Section 3.2, we introduce system models. In Section 3.3, we presents the proposed traffic modeling scheme, investigates the effects of mobility and spatial correlation on the traffic statistics, and

introduces the traffic shaping schemes using movement coordination. Section 3.4, we present experimental results.

3.2 System Models

In this section, we present the characteristics of WSANs including node mobility and spatial correlation, all of which have significant impact on the traffic nature in WSANs.

3.2.1 Node Mobility

Generally, a WSAN has two types of mobile nodes: robot and human agent. Compared with human agent, the movement pattern of robots could exhibit much more regularity and predictability because they are either remotely controlled or preprogrammed. Since human agent can have a wide range of mobility variability to cover more scenarios, the mobility patterns of human agent is of interest in this work. However, as seen in the following sections, the conclusions of this chapter also provide valuable guidelines on how to design the effective resource provisioning strategy for WSNs with any type of mobile agents.

The recent seminal work [44] has investigated the human mobility features based on the real GPS traces. The data reveal that the statistical pattern of human movements can be characterized by a two-state process alternating between pausing and moving. The distance a human object traveled during the moving state is defined as flight. The length of a flight is measured by the longest straight line trip from one location to another that the human object makes without a directional change. The flight length has been revealed to follow heavy-tail distribution [44]. Accordingly, its survival function can be expressed by

$$\overline{F_d}(x) = P(X \geq x) = \left(\frac{b}{x}\right)^{\alpha}, \quad x \geq b \tag{15}$$

where $b \geq 0$ denotes the minimum distance a human agent can travel and α denotes

tail index. According to [44], the tail index will be close to 1 for the outdoor environment. In this case, the human mobility will exhibit high variability since the flight length will drastically fluctuate within a wide range of values over three-orders of magnitudes (i.e., 1000 meters). This high mobility variability has been shown to be determined by human intentions to travel from one position to another without much deviation caused by geographical constraints such as roads and buildings.

3.2.2 Spatial Correlation

Besides node mobility, spatial correlation is another significant characteristic of WSANs. For typical WSAN applications, the mobile sensor nodes are required to observe the interested phenomenon at different locations in the field and send the measured data to the sink(s). The observed phenomenon is usually a spatially dependent continuous process, in which the measured data have a certain spatial correlation. In general, the degree of the spatial correlation in the data increases with the decrease of the separation between two observing locations. To quantify the spatial correlation, the observations $S_1, S_2, ..., S_N$ at N locations are modeled as an N-dimensional random vector $S = [S_1, S_2, ..., S_N]^T$ [8] [50] which has a multivariate normal distribution with $[0, 0, ..., 0]^T$ mean and covariance matrix K with each element defined by

$$k_{ij} = \frac{E[S_i S_j]}{\sigma^2}, \quad i, j = 1, 2, ..., N \tag{16}$$

k_{ij} denotes a correlation function that specifies the correlation model. The correlation function is nonnegative and decreases monotonically with the distance $d_{(i,j)}$ between two locations i and j. Correlation models can be categorized into four groups as Spherical, Power Exponential, Rational Quadratic, and Matern [50]. Each of them characterizes the properties of different physical phenomena

3.3 Spatial Correlation and Mobility Aware Traffic Modeling

In this section, we propose a structural traffic modeling scheme, which aims to mimic the typical behavior of the mobile node. This scheme is favored because it yields a traffic model whose parameters are related to the traffic generating mechanism and the main attributes of the network (e.g., mobility and spatial correlation). Consequently, it could provide insight into the impact of network design parameters and control strategies on the pattern of the generated traffic. In the rest of this section, we first abstract the behavior of the mobile agent, and then the corresponding traffic model is presented. Based on the proposed model, the statistical analysis is given, and the traffic shaping schemes are proposed.

3.3.1 Single Mobile Node Traffic

In a WSAN, the behavior of the mobile sensor node can be described by a procedure having two phases: sensing phase and transmitting phase. During the sensing phase, the node moves to a location, executes sensing tasks, and performs in-network data compression. During the transmitting phase, the compressed data are sent at certain rate using suitable transmission mechanisms. This node behavior implies that the transmission pattern of the node can be naturally characterized by a two-state process that alternates between transmission and silence. According to this abstracted node behavior, we utilize ON/OFF process X(t) to model the single-node traffic, which alternates between two states: the ON state, during which the source transmits data at a rate r; and the OFF state, during which the source is silent. Let $\tau_a(i)$ and $\tau_b(i)$ denote the duration of the ith ON state and the ith OFF state, respectively. The traffic generated by a single sensor node, versus time, can be mathematically modeled as

$$X(t) = \sum_{n=0}^{\infty} r_{[T(n),T(n)+\tau_a(n+1))}(t), t \geq 0, \tag{17}$$

where $T(n)$ denotes the time of occurrence of the nth ON period, i.e.,

$$T(n) = T(0) + \sum_{i=1}^{n} (\tau_a(i) + \tau_b(i)),\ n \geq 1$$

and $r_{[T(n),T(n)+\tau_a(n+1))}$ is the indicator function, which is equal to r only for $t \in [T(n), T(n) + \tau_a(n+1))$.

To completely characterize this model, the distributions of the ON and OFF periods need to be specified. The distribution of OFF period is affected by the specific sensing operation parameters, such as the sensing time and the information processing time. To generalize the analysis, it is assumed that the OFF period follows any survival function, denoted by $\overline{F_{\tau b}}(x)$. The ON state distribution, denoted by $\overline{F_{\tau a}}(x)$, depends on the amount of data transmitted at each sensing location. This quantity is closely related to on the statistical features of mobility and spatial correlation.

Let V denote the file required to be transmitted during an ON period. The length of the ON period τ_a is simply the time to transmit the file using a certain rate r, i.e.,

$$\tau_a = V/r. \tag{18}$$

It is evident that the time for data transmissions depends on the date rates of the specific sensor platforms. For example, the crossbow MICAz nodes equipped with 802.15.4 transceiver modules can achieve up to 250 kps date rate. On the contrary, the crossbow stargate nodes, which leverage 802.11 transceiver modules, can support up to 11Mps date rate. Without loss of generality, we utilize the constant data rate, e.g., r = 1. In this case, the distribution of the ON period length only depends on the distribution of the file size. To obtain the file-size distribution, we first express the file size V in terms of a set of variables related to the network attributes, e.g., spatial correlation and mobility. Then, the probability density function (PDF) of V is derived based on the distributions of these relevant variables.

After the sensor network is deployed, each static sensor node can collect a large

number of sensing samples as time proceeds, and then treats n consecutive sensing samples as an event. To preserve the limited memory spaces of the sensors, each sensor is only required to store the smallest collection of events, which has probability nearly one. More specifically, given a predefined small value δ, each sensor only stores the event e that occurs with probability $P(e)$ satisfying

$$2^{-n(H(S)+\varepsilon)} \leq p(e) \leq 2^{-n(H(S)-\varepsilon)}, \tag{19}$$

where $H(S)$ is the entropy of the observation S. Accordingly, by the asymptotic equipartition property (AEP) theorem, the collection C of such events has probability nearly 1, that is,

$$p(\mathrm{C}) = \sum_{e \in C} p(e) > 1 - \varepsilon. \tag{20}$$

As ε is made arbitrarily small, the total number of events in the collection C, i.e., the cardinality of C, approximates

$$|\mathrm{C}| = 2^{nH(S)}. \tag{21}$$

Obviously, $nH(S)$ bits are sufficient for indexing the events in the collection C. Besides these bits, extra bits are required for conveying more information, such as the node identifications and the location information of the sensors. Therefore, we assume that the constant number B of bits is used to represent each event. Without loss of generality, we let $B = 1$. Accordingly, the total file size for all events in the collection C approximates

$$V = 2^{nH(S)}. \tag{22}$$

After each movement, a mobile agent retrieves the event information from the closest static sensor node. Intuitively, the spatial proximal static sensor nodes trend to detect and record the same events at the same time instances. Consequently, after each movement, a mobile agent needs to collect and report the previously undetected events, i.e., the events that are different from the ones detected at previous location.

More specifically, thanks to the spatial correlation, for a typical event detected at previous location j, at most $2^{nH(S_i|S_j)}$ new events can occur at current location i, where $H(S_i|S_j)$ is the entropy of the observation S_i given the observation S_j. Thus, the number of these new events V reported by the mobile agent at current location is a function of conditional entropy, which is expressed by

$$V = 2^{nh(S_i|S_j)} = 2^{n(h(S_iS_j)-h(S_j))}. \tag{23}$$

Specifically, differential entropy $h(S_i)$ is used instead of discrete entropy $H(S_i)$ because the observed phenomenon S is generally a continuous random process. Note that differential entropy differs from discrete entropy by only a constant if the samples are quantized with an identical and sufficient small quantization step. This constant only affects the resolution of quantization and the packet size, but does not change the resulting amount of samples to be transmitted. The file size V in equation (19) can be evaluated by adopting the power exponential correlation model, which is a commonly used model in the WSN studies due to its capability of characterizing a wide range of phenomena [50]. The correlation function of the adopted model is expressed by

$$k = e^{-(\theta_1 d)^{\theta_2}}, \tag{24}$$

where d is the mutual distance of two locations. The parameters θ_1 and θ_2 control the correlation level within a given distance d. As a general rule of thumb, smaller value of θ_1 or θ_2 indicates higher level of correlation. Based on the above spatial correlation model, the joint entropy of S_i and S_j is given by

$$h(S_iS_j) = \frac{1}{2}\log(2\pi e\sigma^2)^2|K|, \tag{25}$$

where K is covariance matrix and $|K|$ is the determinant of K. That is,

$$K = \begin{bmatrix} 1 & e^{-(\theta_1 d_{ij})^{\theta_2}} \\ e^{-(\theta_1 d_{ij})^{\theta_2}} & 1 \end{bmatrix} \tag{26}$$

Inserting the equation (25) into (19) leads to the closed expression regarding the size of a file conveyed in an ON period

$$V = V_{\max}(1 - e^{-2(\theta_1 d)^{\theta_2}})^{\frac{n}{2}}, \tag{27}$$

where $V_{max} = (2\pi e\sigma^2)^{n/2}$ is the maximum file size, which depends on the properties of the physical phenomenon, e.g., variance σ, and d is the traveled distance or the flight length in the preceding OFF period. Equation (27) shows that for a given phenomenon of interest, the file size distribution depends on the distribution of the flight length defined in (1). Accordingly, the survival function of the ON period length is given by

$$\overline{F_{\tau_a}}(x) = (2^{1/\theta_2} b\theta_1)^{\alpha} (\ln(\frac{x^{2/n}{}_{\max}}{x^{2/n}{}_{\max} - x^{2/n}}))^{-\alpha/\theta_2}, \tag{28}$$

where $x \in [x_{min} \;\; x_{max}]$, $x_{min} = \left(2\pi e\sigma^2(1 - e^{-2(\theta_1 b)^{\theta_2}})\right)^{n/2}$, and $x_{max} = (2\pi e\sigma^2)^{n/2}$.

Assuming the minimum file size of one and the unit transmission rate (r = 1) yields the simplified form of the survival function of the ON period length

$$\overline{F_{\tau_a}}(x) = (\ln \frac{x_{max}^{2/n}}{x_{max}^{2/n} - 1})^{\alpha/\theta_2} (\ln \frac{x_{max}^{2/n}}{x_{max}^{2/n} - x^2})^{-\alpha/\theta_2}. \tag{29}$$

Equation (29) shows that the distribution of the ON period length is determined by two factors: the degree of spatial correlation, e.g., θ_2, and mobility variability, e.g., α. Larger θ_2 or α indicates smaller degree of spatial correlation or smaller mobility variability. To facilitate further analysis, we define characteristic index as

$$\beta = \frac{\alpha}{n\theta_2}. \tag{30}$$

This index β reflects the joint effect of mobility and spatial correlation that has direct impact on the traffic patterns.

3.3.2 Statistical Analysis

In this section, we derive the autocorrelation function of the single node traffic. Particularly, we investigate the inherent relationship between the autocorrelation function

and network attributes, such as mobility and spatial correlation. The revealed result could point us in new directions for designing efficient and effective protocols for WSANs. These protocols are briefly introduced in the next section.

Before evaluating the autocorrelation function, we first investigate the properties of the ON period length, which have a profound impact on the characteristics of the autocorrelation function.

Proposition 1. *If the characteristic index $\beta < 1$, then the length of the ON period follows a power law probability density with tail index $\gamma_a \approx 2\beta < 2$ within the characteristic region $x \in [x_{min}x^*]$, where*

$$x^* = (2\pi e\sigma^2 R^*)^{n/2} \tag{31}$$

and

$$R^* = \arg\{R|\frac{R}{(R-1)\ln(1-R)} = \frac{1}{\beta}\} \tag{32}$$

Proof. According to equation (29), we can express the tail index as

$$\gamma(x) = \frac{d\log\bar{F}(x)}{d\log(x)} = \frac{2\beta x^{2/n}}{x_{\max}^{2/n} - x^{2/n}} / \ln(\frac{x_{\max}^{2/n}}{x_{\max}^{2/n} - x^{2/n}}) \tag{33}$$

Equation (33) indicates that the tail index is a function of time x. It is easy to show that $\gamma(x)$ is an increasing function with respect to x because the derivative of the function $\kappa(x,\beta)$ with respect of x is larger than 0

$$\frac{d\gamma(x)}{dx} = 2x(x_{\max}^{2/n} - x^{2/n})\left(\ln\frac{x_{\max}^{2/n}}{x_{\max}^{2/n} - x^{2/n}}\right)^{-2} \geq 0 \tag{34}$$

Therefore, we can obtain the minimum tail index γ_{min} within the region $x \in [t_{min}\ \ t_{max}]$ by evaluating $\gamma(x)$ at t_{min}. By letting $t = 2\theta_1 b^{\theta_2}$, we have

$$\begin{aligned}\gamma_{min} = \gamma(t_{min}) &= \frac{2\beta(1 - \mathrm{e}^{-(\theta_1 b)^{\theta_2}})}{\mathrm{e}^{-(\theta_1 b)^{\theta_2}}(\theta_1 b)^{\theta_2}} = \frac{2\beta(1-e^{-t})}{e^{-t}t} \\ &\approx \frac{2\beta(1+t-1)}{t} = 2\beta\end{aligned}$$

The above approximation holds when t is small, which is the case in the considered network scenario. More specifically, the value of t is determined by θ_1, θ_2, and b. θ_1 and θ_2 control the correlation degree within a given distance. In general, $\theta_2 \in (0, 2]$ and $\theta_1 < 0.1$ are used to model the spatial correlation of the physical event information [50]. On the other hand, b is the minimum traveled distance of the mobile agent. According to [44], this parameter is generally less than 10 m. As a consequence, $t = 2\theta_1 b^{\theta_2}$ can approach a small value and thus Equation (35) holds.

Therefore, by (35), if $\beta < 1$, then $\gamma_{min} < 2$. Since $\gamma(x)$ is a increasing function, this implies that there exists a region of $x \in [x_{min} \quad x^*]$ with the tail index $\gamma(x) \leq 2$. Therefore, we have the upper bound x^* as follows

$$x^* = \arg\{x|\gamma(x) = 2\} \tag{35}$$

To determine the upper bound t^*, we define $R = (x/x_{max})^{2/n}$ and simplify $\gamma(x)$ in (33) as

$$\gamma(x) = \frac{2\beta R}{(R-1)\ln(1-R)} \tag{36}$$

Let equation (36) be equal to 2. We can obtain R^* as a function of β, i.e.,

$$\mathrm{R}^* = \arg\{R|\frac{R}{(R-1)\ln(1-R)} = \frac{1}{\beta}\}$$

. According to (35), we obtain the upper bound

$$x^* = (2\pi e\sigma^2 R^*)^{n/2} \tag{37}$$

Based on numerical results, it is easy to show R^* decreases as β increases. More specifically, $R^* \approx 0.7$ for $\beta = 0.5$, while R^* approaches 0 as β approaches 1.

□

Remark 5. *Proposition 1 shows that the distribution of the ON period length displays different behavior within different regions. If $x \in [t_{min} \quad t_{max}R]$, then the survival function decays slowly. The decay speed is completely controlled by β or the joint*

effects of spatial correlation degree (α) and mobility variability (θ_2). By contrast, the survival function of the ON period length decays fast if $x > t_{max}R$. This implies that the ON period distribution can be closely approximated by a mixture model.

Proposition 2. *If correlation index $\beta < 1$, the $\overline{F_{\tau_a}}(x)$ can be approximated by mixture Pareto-exponential distribution.*

$$\overline{F^*_{\tau_a}}(x) = \begin{cases} 1 & 0 \leq x < k_1 \\ (k)^{\gamma_a} x^{-\gamma_a} & k_1 \leq x < k_2 \\ \mu e^{-\lambda x} & k_2 \leq x < k_3 \\ 0 & x \geq k_3 \end{cases} \tag{38}$$

where $\gamma_a < 2$. $k_1 = x_{min}$ and $k_3 = x_{max}$ indicate the lower and upper bounds of the possible length of the ON period, respectively. $k_2 = x^*$ is upper bound of the characteristic region defined in Proposition 1.

Proof. Proposition 1 shows the survival function of the ON period length follows power law form with the tail index $\gamma_a = 2\beta$ in the region $t_{min} < x < t_{max}R$. In this case, the survival function can be characterized by Pareto distribution, which is given by

$$\bar{F}^*_{\tau_a}(x) = (\frac{k_1}{x})^{\gamma_a}, \ x \in [k_1 \ k_2] \tag{39}$$

where $k_1 = x_{min}$ and $k_2 = x^*$. In addition, Proposition 1 points out if $x > t_{max}R$, then the tail index is always larger than 2. This means the survival function in this region decays much faster than the heavy tail distribution. Thus, exponential distribution can be used to approximate the survival function within the region $x^* < x < t_{max}$. That is,

$$\bar{F}^*_{\tau_a}(x) = \mu e^{-\lambda x}, \ x \in [k_2 \ k_3] \tag{40}$$

where $k_2 = x_{max}R$ and $k_3 = x_{max}$. The exponent λ can be obtained by the exponent index function of $\bar{F}_{\tau_a}(x)$, which is given by

$$\vartheta(x, \beta) = \frac{d \log \bar{F}_{\tau_a}(x)}{dx} \tag{41}$$

Since the initial point of the exponential distribution is located at $(k_2, \bar{F}_{\tau_a}(k_2))$, then the parameters λ and μ can be given by

$$\lambda = \vartheta(k_2) \tag{42}$$

and

$$\mu = \bar{F}_{\tau_a}(k_2)e^{\vartheta(k_2)k_2} \tag{43}$$

□

Remark 6. *Proposition 2 states that the ON period length follows power law (hyperbolic) form within the region $x_{min} < x < x^*$. Obviously, if x_{max} is large enough, then ON period length approaches heavy tail distribution, which actually characterizes the zero frequency behavior. However, due to the boundary of x_{max}, the middle frequency behavior of the single node traffic is of interest, which is further explained in Proposition 3.*

Based on the approximated survival function defined in (38), we proceed to derive the autocorrelation function of the single node traffic. To express the result in a closed form, we need to specify the survival function of the OFF period. To generalize the analysis, we assume that the OFF period follows Pareto distribution so that OFF period can exhibit a wide range of variability by adjusting the corresponding parameters. The PDF of the OFF period τ_b is expressed by

$$f_{\tau_b}(x) = \gamma_b \frac{m^{\gamma_b}}{x^{(\gamma_b+1)}} \tag{44}$$

where m denotes the minimum OFF time and γ_b denotes the tail index which controls the variability of OFF interval. Values of γ_b smaller than unity induces OFF durations with infinite mean, whereas values of γ_b exceeding two ensure finite variance. In the range of $1 < \gamma_b < 2$, the OFF period lengths have finite mean but exhibit wild variation about that mean as a result of the infinite variance in that range. In sum, the variability of OFF period increases as γ_b decreases. In addition, to simplify the

analysis, we consider the case in which the OFF period is dominated by the time for sensing and information processing, thus implying that the OFF and ON period lengths are independent. To obtain the autocorrelation function, we first investigate the properties of the power spectrum density (PSD).

Proposition 3. *If* $0.5 < \beta < 1$ *and* $\tau_b > 2$*, the power spectrum density* $S(f)$ *of* $X(t)$ *exhibits power law decay behavior with fractal exponent* $2-2\gamma_a$ *in mid-frequency range* $k_2^{-1} \ll f \ll k_1^{-1}$.

Proof. (1) Mid-frequency approximation. Based on the power spectrum density of an ON/OFF process with unit amplitudes and arbitrary distributions of ON and OFF period lengths [31], the power spectrum density of the random process defined by (17) is given by

$$\begin{aligned} S(\omega) = & r^2 E[X(t)]\delta(\frac{\omega}{2\pi}) \\ & + \frac{2r^2(\omega)^{-2}}{E[\tau_a]+E[\tau_b]} \mathrm{Re}\left\{ \frac{[1-\varphi_{\tau_a}(\omega)][1-\varphi_{\tau_b}(\omega)]}{1-\varphi_{\tau_a}(\omega)\varphi_{\tau_b}(\omega)} \right\} \end{aligned} \tag{45}$$

where φ_{τ_b} and φ_{τ_b} are the characteristic functions associated with the distributions for ON period length τ_a and off period length τ_b, respectively. Based on the survival function of τ_a given in equation (38), we can get the characteristic function of τ_a given as

$$\begin{aligned} \varphi_{\tau_a}(\omega) = & \int_{k_1}^{k_2} e^{-j\omega x} \gamma_a \frac{k_1^{\gamma_\alpha}}{x^{\gamma_a+1}} dx \\ & + \int_{k_2}^{k_3} e^{-j\omega x} \mu(\lambda e^{-\lambda x} + \delta(x-k_3)e^{-\lambda k_3}) dx \\ = & \gamma_a (j\omega k_1)^{\gamma_a} \int_{j\omega k_1}^{j\omega k_2} e^{-x} x^{-(\gamma_a+1)} dx \\ & + \frac{\lambda\mu}{j\omega+\lambda} e^{-(j\omega+\lambda)k_2} + \frac{j\omega\mu}{j\omega+\lambda} e^{-(j\omega+\lambda)k_3} \end{aligned} \tag{46}$$

Since $k_3^{-1} < k_2^{-1} \ll f \ll k_1^{-1}$, we have $k_1/k_2 \to 0$, $\omega k_2 \to \infty$, and $\omega k_3 \to \infty$. Let $j\omega k_1 = z$, we can obtain

$$\begin{aligned} 1 - \varphi_{\tau_a}(\omega) &= \gamma_a (j\omega k_1)^{\gamma_a} \int_{j\omega k_1}^{j\omega k_2} (1 - e^{-x}) x^{-(\gamma_a+1)} dx \\ &\quad + (\frac{k_1}{k_2})^{\gamma_a} + \frac{\lambda\mu}{j\omega + \lambda} e^{-(j\omega+\lambda)k_2} \\ &\quad + \frac{j\omega\mu}{j\omega + \lambda} e^{-(j\omega+\lambda)k_3} \\ &\to \gamma_a (j\omega k_1)^{\gamma_a} \int_{j\omega k_1}^{\infty} (1 - e^{-x}) x^{-(\gamma_a+1)} dx \\ &= \gamma_a z^{\gamma_a} \int_z^{\infty} (1 - e^{-x}) x^{-(\gamma_a+1)} dx \end{aligned} \tag{47}$$

Since $0.5 < \beta < 1$, we have $1 < \gamma < 2$. Then, the expansion of above integration around the origin of z is given by [31]:

$$\begin{aligned} Q(x) &= \gamma_a z^{\gamma_a} \int_z^{\infty} (1 - e^{-x}) x^{-(\gamma_a+1)} dx \\ &\to \frac{\gamma_a}{\gamma_a - 1} z - (\gamma_a - 1)^{-1} \Gamma(2 - \gamma_a) z^{\gamma_a} \end{aligned} \tag{48}$$

We therefore obtain the characteristic function of τ_a

$$\varphi_{\tau_a}(\omega) \to 1 - \varsigma_1 j\omega + \varsigma_2 (j\omega)^{\gamma_a} \tag{49}$$

where

$$\varsigma_1 = \frac{k_1 \gamma_a}{\gamma_a - 1}, \ \varsigma_2 = (\gamma_a - 1)^{-1} \Gamma(2 - \gamma_a) k_1^{\gamma_a} \tag{50}$$

Because the OFF period length τ_b follows Pareto distribution with minimum value of m and the tail index $\gamma_b > 0$, following the same procedures as above, we can obtain the characteristic function of τ_b given as

$$\begin{aligned} \varphi_{\tau_b}(\omega) &= \int_m^{\infty} e^{-j\omega x} \gamma_b \frac{m^{\gamma_b}}{x^{\gamma_b+1}} dx \\ &\to 1 - \rho_1 j\omega + \rho_2 (j\omega)^{\gamma_b} \end{aligned} \tag{51}$$

where

$$\rho_1 = \frac{m\gamma_b}{\gamma_b - 1},\ \rho_2 = (\gamma_b - 1)^{-1}\Gamma(2-\gamma_b)m^{\gamma_b} \tag{52}$$

Inserting equations (49) and (51) into (45), the latter part of the second term of (45) becomes

$$\begin{aligned}
\Omega(\omega) &= \mathrm{Re}\left\{\frac{[1-\varphi_{\tau_a}(\omega)][1-\varphi_{\tau_b}(\omega)]}{1-\varphi_{\tau_a}(\omega)\varphi_{\tau_b}(\omega)}\right\} \\
&= \mathrm{Re}\left\{\frac{\varsigma_1\rho_1\omega^2 - \varsigma_2\rho_1(j\omega)^{\gamma_a+1} - \varsigma_1\rho_2(j\omega)^{\gamma_b+1}}{(\varsigma_1+\rho_1)j\omega - \varsigma_2(j\omega)^{\gamma_a} - \rho_2(j\omega)^{\gamma_b}}\right\} \\
&= -\frac{\varsigma_2\rho_1^2\cos(\frac{\pi}{2}\gamma_a)}{(\varsigma_1+\rho_1)^2}\omega^{\gamma_a} - \frac{\rho_2\varsigma_1^2\cos(\frac{\pi}{2}\gamma_b)}{(\varsigma_1+\rho_1)^2}\omega^{\gamma_b}
\end{aligned} \tag{53}$$

Based on the survival function of τ_a given in equation (38), we can obtain the mean of τ_a

$$\begin{aligned}
E[\tau_a] &= \int_{k_1}^{k_2}\gamma_a x\frac{k_1^{\gamma_a}}{x^{\gamma_a+1}}dx \\
&\quad + \int_{k_2}^{k_3} x\mu(\lambda e^{-\lambda x} + \delta(x-k_3)e^{-\lambda k_3})dx \\
&= \frac{\gamma_a}{\gamma_a - 1}(k_1 - k_2(\frac{k_1}{k_2})^{\gamma_a}) \\
&\quad + (k_2\mu + \frac{\mu}{\lambda})e^{-\lambda k_2} - \frac{\mu}{\lambda}e^{-\lambda k_3}
\end{aligned} \tag{54}$$

Based on the survival function of Pareto distribution, we can obtain the mean of τ_b

$$E[\tau_b] = \frac{\gamma_a}{\gamma_a - 1}m \tag{55}$$

Inserting equations (53), (54), and (55) into (45), we can obtain the mid-frequency approximation of the power spectrum density with $k_2^{-1} \ll f \ll k_1^{-1}$

$$S(\omega) = C_1\omega^{\gamma_a - 2} + C_2\omega^{\gamma_b - 2} \tag{56}$$

where

$$C_1 = -\frac{2(\gamma_a-1)^{-1}\Gamma(2-\gamma_a)k_1^{\gamma_a}(\frac{m\gamma_b}{\gamma_b-1})^2 r^2\cos(\frac{\pi}{2}\gamma_a)}{(E[\tau_a]+E[\tau_b])(\frac{k_1\gamma_a}{\gamma_a-1} + \frac{m\gamma_b}{\gamma_b-1})^2} \tag{57}$$

and

$$C_2 = -\frac{2(\gamma_b - 1)^{-1}\Gamma(2-\gamma_b)m^{\gamma_b}(\frac{k_1\gamma_a}{\gamma_a - 1})^2 r^2 \cos(\frac{\pi}{2}\gamma_b)}{(E[\tau_a] + E[\tau_b])(\frac{k_1\gamma_a}{\gamma_a - 1} + \frac{m\gamma_b}{\gamma_b - 1})^2} \tag{58}$$

Because $\tau_b > 2$ and $\tau_a \approx 2\beta < 2$, the spectrum density has a fractal exponent $2 - 2\beta$ in mid-frequency.

(2) High-frequency approximation. If $\omega \to \infty$, we can obtain the characteristic functions of ON and OFF periods, respectively,

$$\varphi_{\tau_a}(\omega) \to 0; \varphi_{\tau_b}(\omega) \to 0$$

Inserting above equations into (45), we can obtain the asymptotic form of the power spectrum density in the high-frequency limit

$$S(\omega) = \frac{2r^2(\omega)^{-2}}{E[\tau_a] + E[\tau_b]} \tag{59}$$

(3) Low-frequency approximation. If $\omega \to 0$, we can obtain the characteristic functions of ON and OFF periods, respectively,

$$\varphi_{\tau_a}(\omega) \to 0; \varphi_{\tau_b}(\omega) \to 1 - \rho_1 j\omega + \rho_2(j\omega)^{\gamma_b}$$

Inserting above equations into (45), we can obtain the asymptotic form of the power spectrum density in the low-frequency limit

$$\begin{aligned} S(\omega) &= \frac{2r^2(\omega)^{-2}}{E[\tau_a] + E[\tau_b]} \mathrm{Re}\left\{1 - \varphi_{\tau_a}(\omega)\right\} \\ &= \frac{-2r^2(\gamma_b - 1)^{-1}\Gamma(2-\gamma_b)m^{\gamma_b}\cos(\frac{\pi}{2}\gamma_b)}{E[\tau_a] + E[\tau_b]} \omega^{\gamma_b - 2} \end{aligned} \tag{60}$$

□

Remark 7. *Proposition 3 states that the spectrum density follows power law form in mid-frequency. According to the relationship between spectrum density and autocorrelation, we can obtain the hyperbolic autocorrelation function within a certain region.*

Proposition 4. *If* $0.5 < \beta < 1$ *and* $\tau_b > 2$*, the autocorrelation function of the process* $X(t)$*, denoted by* $R_m(\tau)$*, follows the power law form*

$$R(\tau) = D_1 t^{1-\gamma_a} + D_2 t^{1-\gamma_b}$$

in the region $x_{\min} \ll |\tau| \ll x_{\max} R$ *with some constants* D_1 *and* D_2*. And the corresponding Hurst parameter is given by*

$$H \approx \frac{3-2\beta}{2}$$

Proof. Based on the relationship between the PSD of a real-valued process and its autocorrelation, we obtain the autocovariance $C_X(t)$ based on the PSD given by (56)

$$\begin{aligned} C_X(t) &= R_X(t) - E^2[X(t)] \\ &= 2\int_{0^+}^{\infty} S(f)\cos(2\pi f t)df \\ &= C_1\pi^{-1}t^{1-\gamma_a}\int_0^{\infty} x^{\gamma_a-2}\cos(x)dx \\ &\quad + C_2\pi^{-1}t^{1-\gamma_b}\int_0^{\infty} x^{\gamma_b-2}\cos(x)dx \\ &= \frac{(\gamma_a-1)^{-1}k_1^{\gamma_a}(\frac{m\gamma_b}{\gamma_b-1})^2 r^2}{(E[\tau_a]+E[\tau_b])(\frac{k_1\gamma_a}{\gamma_a-1}+\frac{m\gamma_b}{\gamma_b-1})^2}t^{1-\gamma_a} \\ &\quad + \frac{(\gamma_b-1)^{-1}m^{\gamma_b}(\frac{k_1\gamma_a}{\gamma_a-1})^2 r^2}{(E[\tau_a]+E[\tau_b])(\frac{k_1\gamma_a}{\gamma_a-1}+\frac{m\gamma_b}{\gamma_b-1})^2}t^{1-\gamma_b} \end{aligned} \tag{61}$$

The power law form of autocovariance (61) implies that the single node traffic has the Hurst parameter

$$H = \frac{3-\min(\gamma_a,\gamma_b)}{2} \approx \frac{3-\min(2\beta,\gamma_b)}{2} \tag{62}$$

Since $\beta < 1$ and $\gamma_b > 2$, we obtain

$$H \approx \frac{3-2\beta}{2} \tag{63}$$

□

Remark 8. *Proposition 4 suggests that the joint effects of mobility and spatial correlation can lead to pseudo LRD traffic, which has power law autocorrelation function with the Hurst parameter up to the cutoff time lag x^*. This result is important because when x_{max} is large enough, pseudo LRD traffic exhibits the similar behavior as the LRD traffic, thus demanding the similar resource allocation approaches. In addition, Proposition 4 shows that index β completely controls the value of the Hurst parameter and thus has direct impact on the burstiness of the traffic from each mobile sensor node. This means that the network traffic can exhibit different degree of hurstiness under different monitored environment (e.g., spatial correlation) and node behavior (e.g., mobility variability).*

3.3.3 Relay Node Traffic Model

Relay Node Traffic Model Because of the nature of ad hoc communication in WSANs, some nodes will act as the relay points responsible for forwarding traffic for other nodes. For certain relay nodes, there may exist several routes passing through them, each route corresponding to a traffic flow originating from other node. Thus, at each relay node, the traffic flow is the aggregation of all the flows that need to be forwarded. The number of flows actually traversing the relay node is changing over time due to the time-varying topology induced by the mobility nature of WSANs. Suppose within a certain time interval T each forwarding flow passes through the relay node with an identical probability of P_t, which is called the traversing probability. Then, the traffic on the relay node, denoted by $S(t)$, becomes the sum of all the active traffic flows during a certain time interval T, i.e.,

$$S(t) = \sum_{n=1}^{N} X_n(t), \tag{64}$$

where $X_n(t), i = 1...N$ denotes the active traffic flows, which are independent identically distributed (i.i.d.). Suppose the maximum number of flows a relay node can

forward is M. Then, it can be shown that N follows binomial distribution. In addition, the autocorrelation function of relay node traffic can display power-law behavior as a result of the independence of the traffic active flows and the properties of iterative mean, which is explained in the following proposition.

Proposition 5. *The traffic at relay node has Hurst parameter $H \approx \frac{3-2\beta}{2}$ in the region $x_{min} \ll t \ll x^*$ with the scaled autocovariance function of the single flow traffic, i.e.,*

$$C_S(\tau) = MP_t C_X(\tau) + M(P_t - {P_t}^2)E^2[X(t)], \tag{65}$$

where $C_S(\tau)$ and $C_X(\tau)$ are the autocovariance function of the relay node traffic $S(t)$ and single node traffic $X(t)$, respectively, and M is the maximum number of flows a relay node can forward, which is determined by system parameters such as bandwidth.

Proof. Taking advantage of the independence of the active flow traffics and the properties of iterative mean, we can obtain

$$\begin{aligned}
R_S(\tau) &= E[\sum_{i=1}^{N} X_n(s) \sum_{j=1}^{N} X_m(s+\tau)] \\
&= E[E[\sum_{i=1}^{N} X_n(s) \sum_{j=1}^{N} X_m(s+\tau)|N]] \\
&= E[\sum_{i=1}^{N} \Bigg\{ E[X_n(s)X_n(s+\tau)|N] \\
&\qquad + \sum_{i\neq j}^{K} E[X_n(s)X_m(s+\tau)|N] \Bigg\}] \\
&= E[NR_X(\tau) + N(N-1)E^2[X(t)]] \\
&= MP_t R_X(\tau) + (M^2 P_f - MP_t^2)E^2[X(t)]
\end{aligned} \tag{66}$$

Based on the relationship between autocorrelation and autocovariance, we obtain

$$\begin{aligned}
C_S(\tau) &= R_S(\tau) - E^2[X(t)] \\
&= MP_t R_X(\tau) + (M^2 P_t - MP_t^2)E^2[X(t)] - E^2[S(t)] \\
&= MP_t C_X(\tau) + M(P_t - {P_t}^2)E^2[X(t)]
\end{aligned} \tag{67}$$

where $C_X(\tau)$ is the autocovariance of a single traffic flow. Equation (67) indicates that the aggregated traffic on the relay node has the scaled autocovariance of the single traffic flow. In addition, since the variance of $S(t)$ is given by

$$Var[S(t)] = MP_t\sigma_X^2 \tag{68}$$

we obtain the autocorrelation function of $S(t)$ which has the some form as single flow traffic $X(t)$

$$\rho_S = \frac{C_S(\tau)}{Var[S(t)]} \approx \frac{C_X(\tau)}{Var[X(t)]}$$

when τ is large. Therefore, the relay node traffic has the same Hurst parameter $H \approx \frac{3-2\beta}{2}$ as the single flow traffic. □

Proposition 5 states that the statistics of relay node traffic closely resemble those of the traffic originating from a single user. Because the aggregated traffic exhibits all of the transitions occurring within each individual traffic flow passing through the relay node, the traffic on the relay node exhibits MP_t times more fluctuations than each of the traversing traffic flow.

3.3.4 Useful Findings and Intuitive Explanations

Based on the statistical analysis given above, we have two major findings. First, higher mobility variability (smaller α) along with smaller spatial correlation (larger θ_2) could lead to more bursty traffic (higher H). Second, the joint effect of the mobility and correlation could lead to non-bursty traffic (e.g., $H = 0.5$) if a certain condition holds (e.g., $\beta = 1$). The intuitive explanation for these findings is that both higher mobility variability and smaller spatial correlation encourage larger file sizes, thereby increasing the tail weight of the size distribution. This in turn leads to more chance to have burst transmissions. Therefore, it can be seen that the traffic burstiness actually arises from the joint effect of node mobility and spatial correlation on the file size, which can be explained as follows. Higher mobility variability indicates

higher probability that a node travels a long distance, which means the observation at current location has little correlation with the previous location. Therefore, at current location the sensor node has to transmit a large data file consisting of almost all the gathered information(e.g., images and videos). On the other hand, small spatial correlation implies that the sensing data retrieved at current location may be completely different from the proximal location. As a result, the sensor node may also need to send a large file of data even through it only travels a short distance. In sum, higher mobility variability along with smaller spatial correlation could lead to higher probability for a node to transmit a large file, thereby resulting in more chance for burst transmissions.

3.3.5 Mobility-aware Traffic Shaping Protocol

The above novel findings provide valuable insights into the design of traffic engineering solutions for WSANs. Instead of passively observing the traffic and designing the resource allocation schemes accordingly, new traffic shaping protocols can be proposed to actively shape the traffic so that the resulting traffic can follow the desired characteristics. Then, the existing traffic management schemes can be directly employed accordingly. In our case, because of the direct connection between the mobility pattern and the traffic attributes, we can effectively reduce the burstiness of traffic flows by adaptively coordinating the movement of the sensor nodes so that the protocols designed for SRD traffic can be employed in more dynamic network environment. More specifically, according to Proposition 4 and 5, the best effort to reduce the traffic burstiness is to let each mobile node move evenly in the sensing area so that the resulting Hurst parameter is minimized by maximizing α. To show the effectiveness of this mobility-aware traffic shaping scheme, we next investigate the queueing performance on the relay node and demonstrate how the proposed scheme simplifies the resource provisioning strategy.

It is easy to prove that if the number of the traversing flows on the relay node is large enough, then the relay node traffic, defined in (64), converges to a fractional Gaussian process. Accordingly, the aggregate cumulative packet traffic follows fractional Brownian motion, which result in Weibull-like asymptotic tail probabilities for queue-length distribution [10], i.e.,

$$P(Q > x) \approx \exp(-\delta b^{2-2H}) \tag{69}$$

where δ is a constant and Q is the steady-state queue length of the relay node. It is shown in [23] that the slowly decaying tail probabilities given in (69) can result in the "buffer ineffectiveness" phenomenon, in which increasing the buffer sizes beyond a certain value results in only a slight decrease in loss rates. The conventional approach to solve this problem is to employ the resource provisioning scheme that exploit the small buffer capacity along with large bandwidth. This scheme can effectively curtail the influence of LRD traffic on the queueing performance. However, for the bandwidth-constraint WSANs, this strategy is infeasible. In this case, selecting the proper traffic shaping protocols to reduce traffic burstiness facilitates the design of effective resource provisioning strategies.

Specifically, the distance the mobile nodes will travel can be controlled to follow any distribution with low variability such as the exponential distribution $F_d(x)$, e.g.,

$$\overline{F_d}(x) = \exp(-\lambda x), \tag{70}$$

where $1/\lambda$ is the average flight length and the survival function of the ON period length with unit transmission rate can be expressed by

$$\overline{F_{\tau_a}}(x) \approx \exp(-\eta x^{\frac{2}{\theta_2}}) \quad x \geq 0, \tag{71}$$

where η is a constant. Equation (71) indicates that the length of the ON period does not follow heavy-tail distribution because it decays faster than exponentially. As shown in [21], if the arrival process is a single ON/OFF source in which the

ON periods are exponentially distributed, then the queue length distribution decays exponentially. This observation implies that the simple buffering scheme for SRD traffic can be effectively applied.

3.4 Performance Evaluation

Validating the ON/OFF model for single node traffic We carry out the experiments in a large-sale simulated sensor network deployed in a large monitoring region. All the settings of this network are configured according to above mentioned hybrid network scenario. The actual traffic traces are first collected from this network. Then, we synthesize traffic traces based on the proposed traffic model and show that the synthesized traffic fits the actual traffic. The actual traffic is generated by a mobile agent, which moves around within an experiment region of 5000*5000. The measurements at each location in this region are generated according to the two-dimensional Gaussian random field with the power exponential correlation coefficient, zero mean and 1 covariance. The static sensor nodes first discretize the continuous measurement by encoding it with one of the 7 symbols. Each symbol corresponds to one of the 7 regions including $(-\infty, -2.5)$, $[-2.5, -1.5)$, $[-1.5, -0.5)$, $[-0.5, 0.5)$, $[0.5, 1.5)$, $[1.5, 2.5)$, and $[2.5, \infty)$. Then, the static sensor node assembles 5 consecutive encoded measurements as an event and store all the events locally. Next, the mobile agent initiates its movement according to the human mobility model, that is, for each movement its actual traveled distance is drawn from the heavy tailed distribution with $\alpha = 0.75$. After each movement, the mobile node retrieves the interested events from the closest static sensor node. In this case, the interested events at current location are the new events that occur at the same time as the events the most frequently detected at the previous location, but have different values or encoded symbols. In Figure 2, we show the volume of the actual traffic measured during each transmission period. Since constant transmission rate is adopted, the traffic volume actually indicates the

length of the transmission period. The corresponding CCDF and tail index estimate of the transmission periods are shown in Figure 3, respectively. The tail index is estimated through linear regression in the log-log CCDF plot. The linearity in this figure indicates that the actual traffic traces are indeed heavy tail distributed. Figure 4 depicts the synthesized traffic using the proposed ON/OFF model. The tail index a of the proposed model is obtained through the tail index estimate of the actual traffic traces. We observe that the outlooks of the two traces are very similar, i.e., they are impulsive. The straight line in the tail index estimate plot in Figure 5 clearly shows that the synthesized traffic is heavy tailed distributed.

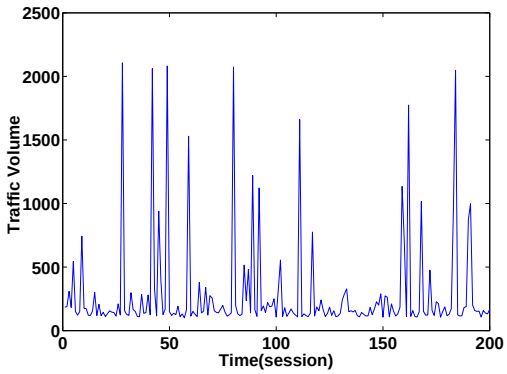

Figure 2: Actual single user traffic.

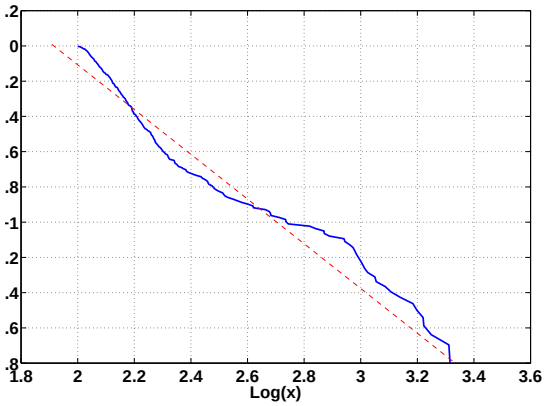

Figure 3: Tail index estimate for actual traffic

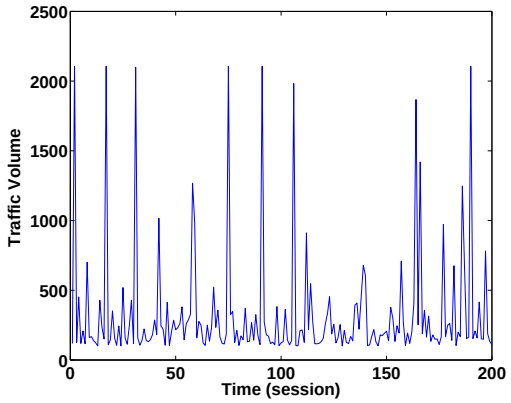

Figure 4: Synthesized single user traffic.

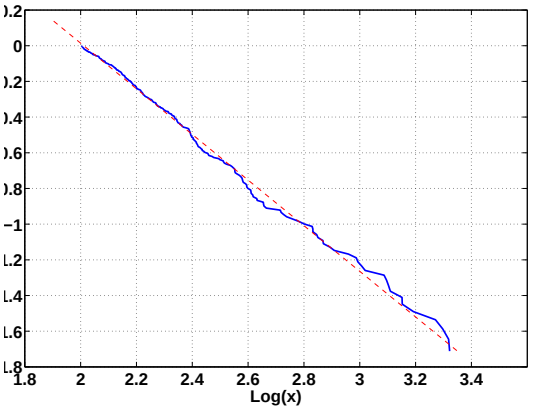

Figure 5: Tail index estimate for synthesized traffic

LRD Behavior in Single Node Traffic Long-range dependence manifests itself either in time-domain with the power-law decay of the autocorrelation or in spectral domain with the power-law singularity of the spectral density at zero frequency. Here, we investigate the corresponding power spectrum density (PSD) in the mid-frequency range because the single node traffic is expected to exhibit pseudo-LRD behavior if the characteristic index $\beta < 1$. Figure 6 gives a rough estimate of PSD of the simulated single user traffic based on the single trace measurement with 10000 sampling points. PSD in Figure 6 shows a trend of power-law decay especially in the low-frequency domain, which indicates the possible presence of long-range dependence. To eliminate the deviation, Figure 7 shows the averaged PSD based on the results from 50 trace measurements and Figure 7 clearly shows the power-law decay of the PSD, which has a straight line with slope 0.8 in the mid-frequency range from one rad to the order of magnitude of 10^{-3} rads. According to the relationship between PSD and autocorrelation, this observation indicates that the corresponding autocorrelation also exhibits power-law behavior. This observation is expected because according to the simulation settings, the correlation parameter β is less than one. Under this condition, Proposition 2 points out that the autocorrelation of the single node traffic obeys a power law in middle frequency, which indicates pseudo-LRD traffic. To determine the Hurst parameter, we employ a periodogram-based analysis. According to this analysis, an estimate of $1 - 2H$ is given by computing the slope of a regression line of the periodogram plotted in the log-log grid. Figure 8 depicts the periodogram of a single trace used in Figure 7. The periodogram plot shows a slope of 0.7376, yielding an estimate of H as 0.8683. The estimated H closely approximates the theoretical $H = 0.9$.

LRD Behavior in Relay Node Traffic Next, we test the long-range dependence of the traffic at relay node. We assume the traversing probability is one, which means

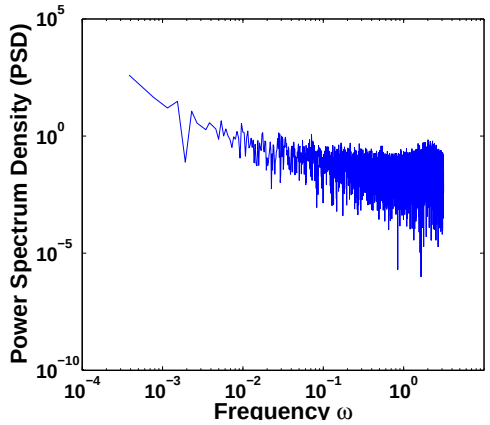

Figure 6: Single node estimated PSD.

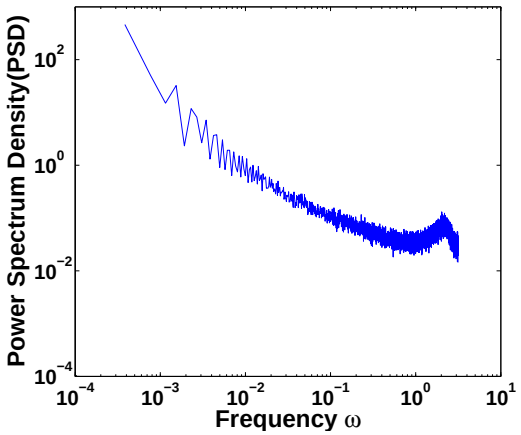

Figure 7: Single node averaged PSD.

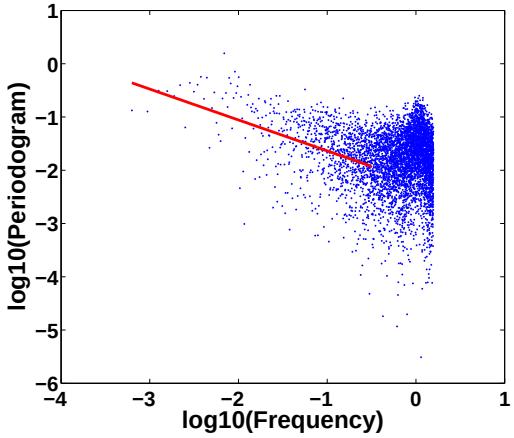

Figure 8: Single node Periodogram.

that each flow in the network is active all the time. The maximum number of flows the relay node can support is set to be 20. Figure 10 compares the aggregated traffic at the relay node with a single flow traffic in terms of PSD. It can be seen that the PSD of the traffic at the relay node closely resembles that of the single flow traffic except that the PSD of the aggregated traffic is 20dB higher than that of the single flow traffic. From the modified periodogram depicted in Figure 9, it can be noticed that the aggregated traffic has the same Hurst parameter as the single flow traffic. These results are expected because according to Proposition 2, the traffic at the relay node has the scaled version of the autocovariance of the single flow traffic. This fact suggests that the PSD of the aggregated traffic is only a shift of the PSD of the single flow traffic in the log-log scale, thus making the Hurst parameter unchanged.

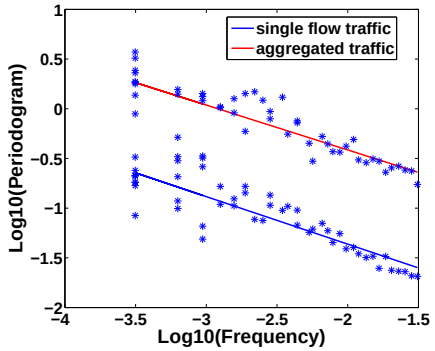

Figure 9: Aggregated traffic vs. single traffic.

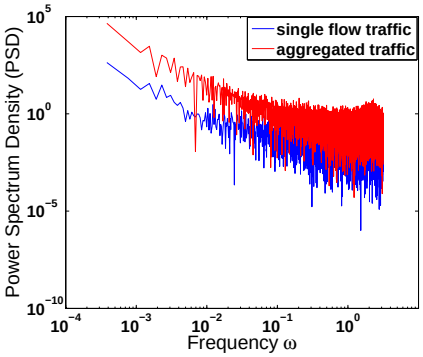

Figure 10: PSD of aggregated traffic vs. PSD of single traffic.

CHAPTER IV

ASYMPTOTIC DELAY ANALYSIS UNDER HEAVY-TAILED ENVIRONMENT

4.1 Introduction

Transmission delay, as one of the key QoS metrics, has been widely studied for classical communication network paradigms the last several decades. Heavy-tailed delay, i.e., the delay following heavy tail distribution exhibits significantly different behavior from that of the light-tailed (e.g., exponential) ones [41]. More specifically, heavy-tailed delay can have infinite moments of lower orders, e.g., mean and variance. In this case, the network can exhibit significant performance degradations including the considerably reduced network throughput, queue stability, and system scalability. Despite its importance, the tail behavior of transmission delay in wireless domain is still an under-explored area, partially due to the dynamic and complex network environment.

In this chapter, we provides an asymptotic analysis of the transmission delay under dynamic spectrum access schemes. Dynamic spectrum access (DSA) is an emerging communication technique that allows the secondary users (SUs) to share the spectrum in an opportunistic manner [2]. Using such scheme, the SUs can access the unoccupied spectrum during idle periods of the primary users (PUs), and stop transmissions when the PU channels become busy. The opportunistic use of the unoccupied spectrum will enable next-generation cellular networks that allow high bandwidth multimedia applications, sensor networks that avoid interference from the coexisting WiFi, military networks that can can be set up in foreign lands, and also allow greater outreach to areas lacking wireless infrastructure

So far, the majority of research in dynamic spectrum access networks focuses on the development of the resource allocation and spectrum management schemes under the assumption of the light tailed behavior of primary and secondary users. In [28] and [61], the queuing delay of SUs in a multi-channel cognitive network was investigated with different objectives. Specifically, using large deviation approximation, [28] aimed to analyze the stationary queue distribution of SUs under the Markov chain based PU traffic model. On the contrary, [61] studied the moments of the SUs' queue length under the PU traffic modeled as an alternating ON/OFF process, where the ON periods follow a general distribution and the OFF periods are exponentially distributed.

Contrary to this conventional light-tailed assumption, significant empirical evidence establishes that both PU and SU traffic can actually exhibit the heavy tailed nature. As for the primary users, it is shown that the call holding time of mobile users in 3G cellular networks and the session duration of licensed users in WLANs show heavy tailed statistics [63][32]. On the other hand, the emerging applications such as mobile internet, multimedia surveillance, video conferencing, and on-line gaming require secondary users to support internet and multimedia traffic, which is inherently bursty and exhibits heavy tailed nature. In spite of its importance, the performance limits of DSA network in the presence of the heavy tailed traffic is still an under-explored area, which, however, can fundamentally challenge the applicability and effectiveness of DSA scheme. For example, recent research shows that such heavy tailed behavior has a significant impact on the performance of spectrum sensing [63], which is the building block for many wireless communication schemes.

We consider a dynamic spectrum access network in which an SU can exploit the spectrum holes of multiple stochastically independent channels. A PU channel is modeled by an alternating renewal process, which alternates between busy periods $\{B_i\}_{i\geq 1}$ and idle periods $\{I_i\}_{i\geq 1}$. An SU is only allowed to transmit during the idle

periods, and avoid transmissions when the PU channels become busy. Upon the arrival of a message with size $L > 0$, the SU first splits it into multiple packets with constant size $L_p > 0$, which are then sent consecutively over PU channels. Accordingly, the total time an SU takes to complete the transmissions of a message is defined as the transmission delay. Apparently, under such generic settings, the transmission delay has a close relationship with the SU message size as well as the PU channel availability. The distributions of the message size and PU busy time can be either heavy tailed (HT) or light tailed (LT), depending on the underlying communication systems and the applications the SUs and PUs demand for. For example, in the earlier 2G voice-oriented cellular systems, empirical measurements show that the call holding times are light tailed, or more specifically, exponentially distributed [46]. On the contrary, heavy tailed distributions have been widely observed in current data-oriented communication networks. For example, the file size on the Internet servers, the web access pattern, and the scene length distribution of VBR (variable bit rate) and MEPG video streams have shown HT statistical characteristics [41]. Moreover, recent empirical evidence shows that the call holding time or channel occupancy time in 3G cellular networks also exhibits the HT nature [62][63].

In this work, we first investigate the delay performance when only a single PU channel is utilized [57] [54]. Specifically, it is shown that DSA induces only light-tailed delay as long as both the busy time of PU channels and the message size of SUs are light-tailed. On the contrary, if either the busy time or the message size is heavy tailed, then the SUs' transmission delay is heavy tailed. For this case, we prove that if one of the busy time or the message size is light-tailed and the other is regularly varying with index α, then the transmission delay is regularly varying with the same index α. As a consequence, the delay has an infinite variance provided $\alpha < 2$ and an infinite mean provided $\alpha < 1$. This implies that SUs can experience extremely high delay variations and even stochastically zero throughout when transmitting messages

with finite mean size. Furthermore, if both the busy time and the message size are regularly varying with index α and β, respectively, then the tail distribution of the delay is as heavy as the one with the smaller index.

Moreover, we investigate the benefits of exploiting the transmission opportunities on multiple PU channels [57] [54]. More specifically, we consider two multiple channel access schemes, namely, spectrum mobility and multi-radio diversity. Under spectrum mobility, if a PU appears in a channel currently used by an SU, the SU vacates the channel immediately and continues its transmission in another idle channel [2]. Under multi-radio diversity, an SU is equipped with multiple radio interfaces so that it can simultaneously access multiple channels. We show that compared with the case in which only a single channel is used, both spectrum mobility and multi-radio diversity can mitigate the degree of heavy tailed delay by increasing the orders of its finite moments.

It is worth to notice that a different application that is related to our work is file fragmentation [37]. In this problem, files are partioned into fragments and transferred over wireless channels. The objective is to find the optimal fragmentation policies that minimize the mean transmission time. Different from the file fragmentation application, in which only one file fragment is sent each time the wireless channel is available, SUs will keep sending packets back-to-back as long as the PU channel is detected as idle. Moreover, in the file fragmentation problem, the channel busy time is assumed to be zero [37]. This assumption is not valid in dynamic spectrum access networks due to the existence of PU activities. In particular, recent work, which is based on real-life measurement data, has identified the heavy tailed behavior in the busy periods of PU channels [63]. This behavior was further shown to have a significant impact on the sensing performance of SUs. However, [63] did not answer how this heavy-tailed behavior of PU channels affects the delay performance of SUs, which is one of the key research problems addressed in this chapter.

The rest of this chapter is organized as follows. Section 4.2 introduces system model. Section 4.3 presents the main results regarding the delay performance of SUs. The impact of spectrum mobility and multi-radio diversity is studied in Section 4.4. The simulation results are presented in Section 4.5.

4.2 System Model

Consider a PU channel and an SU which transmits when the PU channel is idle. Without loss of generality, we assume that the PU channel is of unit capacity. This channel is modeled by an alternating renewal process, which alternates between busy periods $\{B_i\}_{i\geq 1}$ and idle periods $\{I_i\}_{i\geq 1}$. $\{B_i\}_{i\geq 1}$ and $\{I_i\}_{i\geq 1}$ are mutually independent random sequences of i.i.d. random variables with distribution F_B and F_I, respectively. Let $L > 0$ denote the size of the messages generated by the SU, and L is a random variable (r.v.) independent of $\{B_i\}_{i\geq 1}$ and $\{I_i\}_{i\geq 1}$. For each message, the SU divides it into packets with constant size $L_p > 0$, which are then sent over the PU channel. In each idle period I_i, the SU attempts to transmit, and if $I_i > L_p$, the SU sends packets consecutively until the remaining time of the idle period I_i is less than the packet size L_p. Otherwise, if $I_i < L_p$, the SU transmits unsuccessfully and waits for the next idle period for retransmission. An illustration of this model is given in Figure 11.

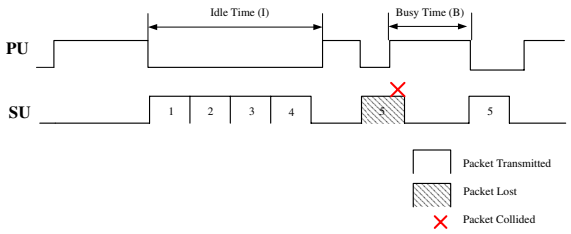

Figure 11: System model.

Definition 6. *During an idle period I_i, the transmission time X_i of the SU is defined*

as

$$X_i := \sup\{nL_p : nL_p \leq I_i\}, \tag{72}$$

the total number of idle periods the SU occupies for transmitting a message of size L is defined as

$$M := \inf\left\{m : \sum_{i=1}^{m} X_i \geq L\right\}, \tag{73}$$

and the total delay T of the SU transmitting a message of size L is defined as

$$T(L) := \sum_{i=1}^{M} \{I_i + B_i\}. \tag{74}$$

4.3 Asymptotic Analysis of the Transmission Delay

In this section, we study the tail asymptotics for the transmission delay experienced by SUs with PU idle times $\{I_i\}_{i\geq 1}$ following LT distribution.

Theorem 1. *If the message size L is heavy tailed, then the number M of idle periods for sending such file is heavy tailed.*

Theorem 2. *If either the busy period B_i or the message size L is heavy tailed, then the transmission delay $T(L)$ is heavy tailed.*

Theorem 3. *If both the busy period B_i and the message size L are light tailed, then the transmission delay $T(L)$ is light tailed.*

Remark 9. *From these results, we see that under the DSA paradigm, SUs can experience light tailed transmission delay if and only if both the message size of SUs and the busy time of PUs are light tailed. In other words, the heavy tailed delay originates not only from the heavy tailed file size but also from the heavy tailed busy time. In this case, the SUs' transmission delay probably has infinite moments of certain orders, e.g., mean and variance, and definitely has an infinite moment generating function, i.e., infinite exponential moments of all orders.*

Proof of Theorem 1. From (73), we have

$$P(M > t) = P\left(L > \sum_{i=1}^{t} X_i\right). \tag{75}$$

Let $\mu := E[X_1]$. For $\varepsilon \in (0, \mu)$, by the law of large numbers, we obtain

$$\begin{aligned}
&P(M > t) = P\left(L > \sum_{i=1}^{t} X_i\right) \\
&\geq P\left(L > \sum_{i=1}^{t} X_i \wedge t(\mu - \varepsilon) < \sum_{i=1}^{t} X_i < t(\mu + \varepsilon)\right) \\
&\geq P(L > t(\mu + \varepsilon)) P\left(t(\mu - \varepsilon) < \sum_{i=1}^{t} X_i < t(\mu + \varepsilon)\right) \\
&\sim P(L > t(\mu + \varepsilon))
\end{aligned}$$

Letting $\varepsilon \downarrow 0$ yields $P(M > t) \gtrsim P(L > \mu t)$. Let $t' = \mu t$. For any $\theta > 0$,

$$\begin{aligned}
\lim_{t\to\infty} e^{\theta t} P(M > t) &\geq \lim_{t\to\infty} e^{\theta t} P(L > \mu t) \\
&= \lim_{t'\to\infty} e^{\frac{\theta}{\mu} t'} P(L > t') \\
&= \infty.
\end{aligned}$$

The last equation holds since L is HT. Thus, M is HT by the Definition 1. □

Proof of Theorem 2. We first consider the case where L is HT. For any $\delta > 0$, we have

$$\begin{aligned}
P(T(L) > t) &= P\left(\sum_{i=1}^{M} I_i + B_i > t\right) \\
&\geq P\left(M \geq \frac{t(1+\delta)}{E[X_1]}\right) \\
&\quad - P\left(\sum_{i=1}^{M} I_i < t \wedge M \geq \frac{t(1+\delta)}{E[X_1]}\right) \\
&\geq P\left(M \geq \frac{t(1+\delta)}{E[X_1]}\right) - P\left(\sum_{i=1}^{t(1+\delta)/E[X_1]} I_i < t\right)
\end{aligned}$$

Let $\widetilde{I}_i := E[I_1] - I_i$. Since I_i is LT, then $\widetilde{I}_i$ is LT. Thus, by applying Chernoff bound, we can argue that there exists a positive constant λ such that for large enough t

$$P\left(\sum_{i=1}^{t(1+\delta)/E[X_1]} I_i < t\right) = P\left(\sum_{i=1}^{t(1+\delta)/E[X_1]} \widetilde{I}_i > \delta t\right) < e^{-\lambda t}.$$

Since L is HT, then M is HT. Thus, for any $0 < \theta < \lambda$

$$\lim_{t\to\infty} e^{\theta t} P(T(t) > t) = \infty. \tag{76}$$

For any $\theta > \lambda$, there always exists a constant $0 < \widetilde{\theta} < \lambda$ such that

$$\lim_{t\to\infty} e^{\theta t} P(T(t) > t) > \lim_{t\to\infty} e^{\widetilde{\theta} t} P(T(t) > t) = \infty. \tag{77}$$

Combining (76) and (77), we have for any $\theta > 0$,

$$\lim_{t\to\infty} e^{\theta t} P(T(t) > t) = \infty. \tag{78}$$

This implies that $T(L)$ is HT by Definition 2.

We will next consider the case where B_i is HT. Since we prove in the previous case that if L is HT, then $T(L)$ is HT, we assume that L is LT. It is easy to see

$$P(T(L) > t) = P\left(\sum_{i=1}^{M} I_i + B_i > t\right) \geq P\left(\sum_{i=1}^{M} B_i > t\right). \tag{79}$$

which implies $T(L)$ is HT provided one can prove $Z := \sum_{i=1}^{M} B_i$ is HT. Towards this, by the independence between M and B_i, we obtain the moment generating function $M_Z(x)$ of Z, i.e.,

$$M_Z(x) = E\left[e^{x\sum_{i=1}^{M} B_i}\right] = E\left[(E[e^{B_1}])^{xM}\right]$$

Since function $f(y) = a^y$ is convex and B_1 is HT, by Jensen's inequality [45], for all $x > 0$

$$M_Z(x) = E\left[(E[e^{B_1}])^{xM}\right] \geq \left(E[e^{B_1}]\right)^{xE[M]} = \infty$$

Thus, it follows that $T(L)$ is HT by the Definition. □

Proof of Theorem 3. By Definition (72), we have $X_i = N_i L_p$ with N as a positive integer random variable, where

$$N_i = \sup \left\{ n : \sum_{i=1}^{n} nL_p \leq I_i \right\}. \tag{80}$$

It is easy to see

$$P(N_i > n) = P(I_i \geq (n+1)L_p) \tag{81}$$

This implies that N_i is LT by Lemma 7. Accordingly, it follows easily from Definition 3 that $X_i = N_i L_p$ is LT. Therefore, invoking Lemma 10(3), we obtain that M is LT. Since $I_i + B_i$ is LT by Lemma 10(1), we finally obtain that $T(L)$ is LT using Lemma 10(2). □

The above Theorems state the conditions under which the SUs' transmission delay exhibits heavy tailed behavior. The following Theorems present the exact asymptotic results for this delay under the regularly varying busy time of PUs and message size of SUs.

Theorem 4. *If $L \in \mathcal{RV}(\alpha)$, then $M \in \mathcal{RV}(\alpha)$ and*

$$P\left(M > t\right) \sim P(L > E[X_1]t). \tag{82}$$

Remark 10. *Comparing with Theorem 1, Theorem 4 provides more refined results regarding the total number of idle periods an SU occupies to transmit a message. Specifically, if the message size is regularly varying, then the number of idle periods for transmitting such message is also regularly varying with the same index. This implies that if the message size has infinite mean and variance, so does the number of idle periods occupied by SUs.*

Theorem 5. *If $L \in \mathcal{RV}(\alpha_l)$ and B_i is LT, then $T(L) \in \mathcal{RV}(\alpha)$ and*

$$P(T(L) > t) \sim P\left(L > \frac{E[X_1]}{E[I_1] + E[B_1]} t\right). \tag{83}$$

Theorem 6. *If $B_i \in \mathcal{RV}(\alpha)$ and L is LT, then $T(L) \in \mathcal{RV}(\alpha)$ and*

$$P(T(L) > t) \sim E[M]P(B_1 > t). \tag{84}$$

Remark 11. *The preceding results establish the relationship between the tail asymptotics of L, B_i, and $T(L)$. Specifically, if one of either the busy time or message size is light tailed and the other is regularly varying, then the tail distribution of the transmission delay is asymptotically proportional to the one with regularly varying distribution. This result implies that SUs can experience extremely high delay variance and stochastically zero throughput even when the transmitting messages are of finite mean size. For example, if the message size is LT, then its mean is finite. In this case, by Theorem 6, when $2 > \alpha > 1$, the transmission delay does not have finite variance, and when $1 > \alpha > 0$, it does not have finite mean, which implies approximately zero throughput on the average.*

Theorem 7. *Assume that $B \in \mathcal{RV}(\alpha_b)$, $L \in \mathcal{RV}(\alpha_l)$, and $E[L] < \infty$. Then, we have*

1. *If $\alpha_b < \alpha_l$, then $T(L) \in \mathcal{RV}(\alpha_b)$ and*

$$P(T(L) > t) \sim E[M]P(B_1 > t) \tag{85}$$

.

2. *If $\alpha_b \geq \alpha_l$,*

$$\lim_{t\to\infty} \frac{\log\left[P(T(L) > t)\right]}{\log t} = -\alpha_l. \tag{86}$$

Corollary 1. *If $B \in \mathcal{RV}(\alpha_b)$, $L \in \mathcal{RV}(\alpha_l)$, and $E[L] < \infty$, then*

$$\lim_{t\to\infty} \frac{\log\left[P(T(L) > t)\right]}{\log t} = -\min(\alpha_b, \alpha_l),$$

and accordingly, the moments of orders $m > \min(\alpha_b, \alpha_l)$ is unbounded, i.e.,

$$E[T(L)^m] = \infty$$

Remark 12. *Comparing the above Theorem and Theorem 4 - 6, we observe that the exact asymptotic tail for the transmission delay is not available in the case of* $\alpha_b \geq \alpha_l$. *Instead, Corollary 1 states that if both busy time and message size are regularly varying, then the tail heaviness of the transmission delay is asymptotically equivalent to the one with smaller index. In this case, it follows directly from Theorem 2 in [11] that the transmission delay still has infinite moments of orders larger than the index* $\min(\alpha_b, \alpha_l)$, *even through this delay does not strictly follow regular varying distributions.*

The proof of Theorem 4 relies on Lemma 11, which we state and prove first.

Lemma 11. *Let* $\widetilde{T}(L) = \sum_{i=1}^{M} I_i$. *If* $L \in \mathcal{RV}(\alpha)$, *then*

$$P(\widetilde{T}(L) > t) \sim P(L > \delta t), \tag{87}$$

where $\delta = E[X_1]/E[I_1]$.

The proof of Lemma 11 relies on Theorem 8 [17]. This technique is similar to the one used in the proof for optimal file fragmentation [37].

Theorem 8. *[17] Let* $L \in \mathcal{RV}(\alpha)$. *Let* $R(t)$ *be a non-negative, almost surely non-decreasing random process independent of* L. *If* $R(t)$ *satisfies following conditions:*

1. $R(t)/t \to \gamma$ *almost surely as* t *goes to infinity, with* $0 < \gamma < 1$.
2. *There exists a positive and finite constant* K *such that* $P(R(t)/t < K) = o(P(L > t))$.

Then $P(L > R(t)) \sim P(L > \gamma t)$

Proof of Lemma 11. We define $N_t := \sup\{n : \sum_{i=1}^{n} I_i < t\}$ and $R(t) := \sum_{i=1}^{N_t} X_i$. It is easy to see that $P(\widetilde{T}(L) > t) = P(L > R(t))$. Thus, to prove Lemma 11, it is

sufficient to prove condition 1 and 2 of Theorem 8 are satisfied. By renewal theory, we have

$$\lim_{t\to\infty}\frac{R(t)}{t}=\frac{E[X_1]}{E[I_1]}=\gamma \tag{88}$$

almost surely. Since $X_1 \leq_{a.s.} I_1$, we conclude $E[X_1] < E[I_1]$ and $0 < \gamma < 1$, implying that condition 1 of Theorem 8 is satisfied. Next, we will prove that condition 2 of Theorem 8 is also satisfied. Let $K = (1-\delta)E[X_1]/((1+\delta)E[I_1])$. Then, for any $1 > \delta > 0$, we have

$$\begin{aligned}P(R(t) < Kt) \leq & P\left(N_t < \frac{t(1-\delta)}{E[I_1]}\right)\\ & +P\left(\sum_{i=1}^{N_t} X_i < Kt \wedge N_t > \frac{t(1-\delta)}{E[I_1]}\right)\\ :=& J_1 + J_2\end{aligned}$$

For J_1, since I_i is LT, by Chernoff bound, there exists $\lambda_1 > 0$ such that

$$J_1 \leq P\left(\sum_{i=1}^{t(1-\delta)/E[I_1]} I_i > t\right) \leq e^{-\lambda_1 t}. \tag{89}$$

For J_2, let $\widetilde{X}_i = E(X_i) - X_i$, we obtain

$$\begin{aligned}J_2 &< P\left(\sum_{i=1}^{t(1-\delta)/E[I_1]} X_i < Kt\right)\\ &= P\left(\sum_{i=1}^{t(1-\delta)/E[I_1]} \widetilde{X}_i > \frac{\delta(1-\delta)E[X_1]}{(1+\delta)E(I_1)}t\right)\end{aligned}$$

Since X_i is LT, by Chernoff bound, we can always find $\lambda_2 > 0$ such that

$$J_2 \leq e^{-\lambda_2 t}. \tag{90}$$

By (89) and (90), we conclude

$$P(R(t) < Kt) \leq e^{-\lambda_1 t} + e^{-\lambda_2 t} \tag{91}$$

Since $L \in \mathcal{RV}(\alpha)$, we have

$$\limsup_{t\to\infty}\frac{P(R(t) < Kt)}{P(L > t)} \leq \limsup_{t\to\infty}\frac{e^{-\lambda_1 t} + e^{-\lambda_2 t}}{t^{-\alpha}\mathcal{L}(t)} = 0 \tag{92}$$

The last equality holds since regularly varying distributions are a subclass of HT distributions. Accordingly, for any $\lambda > 0$, $\lim_{t\to\infty}(e^{\lambda t}t^{-\alpha}\mathcal{L}(t)) = \infty$. By (92), we conclude $P(R(t) < Kt) = o(P(L > t))$. Therefore, both condition 1 and 2 of Theorem 8 are satisfied. This completes tbe proof. □

Proof of Theorem 4. Let $\widetilde{T}(L) = \sum_{i=1}^{M} I_i$. By Lemma 11, (82) follows provided one can show that

$$P(\widetilde{T}(L) > t) \sim P(M > \frac{t}{E[I_1]}) \tag{93}$$

For all $1 > \delta > 0$, we obtain

$$\begin{aligned} P(\widetilde{T}(L) > t) \leq & P\left(M \geq \frac{t(1-\delta)}{E[X_1]}\right) \\ & + P\left(\sum_{i=1}^{M} I_i > t \wedge M \leq \frac{t(1-\delta)}{E[I_1]}\right) \\ \leq & P\left(M \geq \frac{t(1-\delta)}{E[X_1]}\right) + P\left(\sum_{i=1}^{t(1-\delta)/E[X_1]} I_i > t\right) \\ \sim & P\left(M \geq \frac{t(1-\delta)}{E[X_1]}\right) \end{aligned} \tag{94}$$

The last step follows from Chernoff bounds. Letting $\delta \downarrow 0$, this proves the upper bound in (93). As to the lower bound, for all $\delta > 0$, letting $\widetilde{I}_i := E[I_1] - I_i$ yeilds

$$\begin{aligned} P(\widetilde{T}(L) > t) \geq & P\left(M \geq \frac{t(1+\delta)}{E[X_1]}\right) \\ & - P\left(\sum_{i=1}^{M} I_i < t \wedge M \geq \frac{t(1+\delta)}{E[I_1]}\right) \\ \geq & P\left(M \geq \frac{t(1+\delta)}{E[X_1]}\right) - P\left(\sum_{i=1}^{t(1+\delta)/E[X_1]} \widetilde{I}_i > \delta t\right) \\ \sim & P\left(M \geq \frac{t(1+\delta)}{E[X_1]}\right) \end{aligned} \tag{95}$$

Letting $\delta \downarrow 0$, this proves the lower bound in (93). By (94) and (95), we obtain

$$P(M > \frac{t}{E[I_1]}) \sim P(\widetilde{T}(L) > t) \sim P\left(L > \frac{E[X_1]}{E[I_1]}t\right),$$

which implies $P(M > t) \sim P(L > E[X_1]t)$. This completes the proof of (82) and implies $M \in \mathcal{RV}(\alpha)$ by Lemma 5. □

Proof of Theorem 5. The proof follows easily by the similar arguments used in proving Lemma 11. □

To facilitate the proofs of Theorem 6 - 7, we define $T_I := \sum_{i=1}^{M} I_i$ and $T_B := \sum_{i=1}^{M} B_i$. This implies $T(L) = T_I + T_B$.

Proof of Theorem 6. To prove Theorem 6, we first show that T_I is LT and $P(T_B > t) \sim E[M]P(L > t)$. First, we argue that T_I is LT. Since L is LT, from Lemma 10(3), we conclude that M is LT. This implies that $T_I := \sum_{i=1}^{M} I_i$ is LT using Lemma 10(2).

We now show that $P(T_B > t) \sim E[M]P(L > t)$. Since M is independent of B_i and $B_i \in \mathcal{RV}(\alpha)$, it follows that $P(M > t) = o(P(B_i > t))$ invoking Lemma 6. From Lemma 8(1), we see that

$$P(T_B > t) \sim E[M]P(B_1 > t) \tag{96}$$

which, in turn, implies $T_B \in \mathcal{RV}(\alpha_l)$ by invoking Lemma 2(3). We are now ready to prove the upper bound in (84). For any $0 < \delta < 1$

$$\begin{aligned} P(T(L) > t) &= P(T_I + T_B > t) \\ &\leq P(T_B > (1-\delta)t) + P(T_I > \delta t) \\ &\sim P(T_B > (1-\delta)t) \end{aligned}$$

The last step follows since $P(T_I > \delta t) = o(P(T_B > (1-\delta)t))$ using Lemma 6. Letting $\delta \downarrow 0$, this proves the upper bound in (84). As to the lower bound, it is easy to see

$$P(T(L) > t) = P(T_I + T_B > t) \geq P(T_B > t)$$

which, combining with the upper bound, completes the proof of (84). Moreover, (84) implies $T(L) \in \mathcal{RV}(\alpha)$ using Lemma 2(3). This completes the proof. □

Proof of Theorem 7. We first consider the case where $\alpha_b < \alpha_l$. Since $L \in \mathcal{RV}(\alpha_l)$ and $E[M] < \infty$, using Theorem 4, we obtain that $M \in \mathcal{RV}(\alpha_l)$ and $E[M] < \infty$. This, combining with $\alpha_b < \alpha_l$, implies that $P(M > t) = o(P(B_1 > t))$ using Lemma 4. Invoking Lemma 8(1), we conclude that

$$T_B = \sum_{i=1}^{M} B_i \sim E[M]P(B_1 > t).$$

which in turn implies that $T_B \in \mathcal{RV}(\alpha_b)$ by Lemma 5. By Lemma 11, we can see that $T_I \in \mathcal{RV}(\alpha_l)$ since $L \in \mathcal{RV}(\alpha_l)$.

We are now ready to prove the upper bound in (85). For any $1 > \delta > 0$, we obtain that

$$P(T(L) > t) \leq P(T_B > (1-\delta)t) + P(T_I > \delta t) \tag{97}$$

Since $T_I \in \mathcal{RV}(\alpha_l)$, $T_B \in \mathcal{RV}(\alpha_b)$, and $\alpha_b < \alpha_l$, using Lemma 4, we obtain that $P(T_I > \delta t) = o(P(T_B > (1-\delta)t))$. This implies that $P(T(L) > t) \lesssim P(T_B > (1-\delta)t)$ from (97). Letting $\delta \downarrow 0$, we verify the upper bound in (85). As to the lower bound, it is easy to see that $P(T(L) > t) \geq P(T_B > t)$. Since the lower and upper bounds coincide, this completes the proof of (85).

We will next consider the case where $\alpha_b \geq \alpha_l$. Since $L \in \mathcal{RV}(\alpha_l)$, from Lemma 11 and regular variations, we obtain that

$$P(T_I > t) \sim \left(\frac{E[I_1]}{E[X_1]}\right)^{\alpha_l} P(L > t). \tag{98}$$

From Theorem 4, we conclude that $M \in \mathcal{RV}(\alpha_l)$ and

$$P(M > t) = P\left(L > E[X_1]t\right).$$

If $\alpha_b > \alpha_l$, it follows that $P(M > t) = o(P(B_i > t))$. This implies, using Lemma 8(2), that

$$P(T_B > t) \sim \left(\frac{E[B_1]}{E[X_1]}\right)^{\alpha_l} P(L > t). \tag{99}$$

If $\alpha_b = \alpha_l$, using Lemma 9, we obtain that

$$P(T_B > t) \sim E[M]P(B_1 > t) + (E[B_1])^{\alpha_l} P(L > t). \tag{100}$$

Combining (97), (98), (99) and (100), we obtain that

$$\limsup_{t\to\infty} \frac{\log[P(T(L) > t)]}{\log t} \leq -\alpha_l$$

which, in conjunction with

$$\liminf_{t\to\infty} \frac{\log[P(T(L) > t)]}{\log t} \geq \liminf_{t\to\infty} \frac{\log[P(T_I > t)]}{\log t} \geq -\alpha_l$$

completes the proof. □

All the above theorems consider the case where the idle periods $\{I_i\}_{i\geq 1}$ are LT r.v.s. The following Theorem computes the logarithmic asymptotics for the delay under regularly varying idle periods, i.e., $\{I_i\}_{i\geq 1} \in \mathcal{RV}(\alpha_I)$.

Theorem 9. *Assume that $B_i \in \mathcal{RV}(\alpha_b)$. If $L \in$ LT or $L \in \mathcal{RV}(\alpha_l)$ with $\alpha_b < \alpha_l$, we have*

$$\lim_{t\to\infty} \frac{\log\left[P(T(L) > t)\right]}{\log t} = -\alpha_b. \tag{101}$$

Assume that $L \in \mathcal{RV}(\alpha_l)$ and $E[L] < \infty$. If $B_i \in$ LT or $B_i \in \mathcal{RV}(\alpha_l)$ with $\alpha_l \leq \alpha_b$, we have

$$\lim_{t\to\infty} \frac{\log\left[P(T(L) > t)\right]}{\log t} = -\alpha_l. \tag{102}$$

Remark 13. *From the above results, we can see that the tail heaviness (i.e., logarithmic decaying rate) of the delay distribution only depends on either the message size or the busy period, whichever has the heavier tail distribution or the smaller decaying rate, i.e.,* $\min(\alpha_l, \alpha_b)$. *This is consistent with the conclusions made in the case where idle periods are LT r,v.s. This implies that the tail behavior of the idle period distribution has no impact on the tail heaviness of the delay distribution.*

Proof of Theorem 9. The proof follows the similar arguments used in proving the asymptotic results under the case where idle periods are LT r.v.s. □

4.4 Impact of Spectrum Mobility and Multi-radio Diversity

In this section, we study the impact of spectrum mobility and multi-radio diversity on the delay performance of SUs. By spectrum mobility, we mean that if a PU appears in a channel currently used by an SU, the SU should vacate the channel immediately and continue its transmission in another idle channel. By multi-radio diversity, we mean that an SU is equipped with multiple radio interfaces so that it can simultaneously access multiple channels.

Assume that there exist $K \geq 1$ PU channels, which are modeled by K independent alternating renewal processes as defined in section II. Each channel $K \geq j \geq 1$ is denoted by $CH^j = \{(B_i^{(j)}, I_i^{(j)})\})_{i\geq 1}$ and channels $\{CH^j\}_{K\geq j\geq 1}$ are heterogenous, i.e., $\{B_1^{(j)}\}_{K\geq j\geq 1}$ (or/and $\{I_1^{(j)}\}_{K\geq j\geq 1}$) are not identically distributed. To simplify the analysis, we assume that the idle periods are light tailed.

4.4.1 Spectrum Mobility

By spectrum mobility, an SU can switch to the idle channels when its current operating channel is occupied by a PU. As a consequence, the SU sees K channels as a single virtual channel, which stays idle if one of K channels is idle and stays busy if all K channels are busy. This virtual channel can be modeled by a random process that alternates between busy $\{B_i^s\}_{i\geq 1}$ and idle $\{I_i^s\}_{i\geq 1}$ periods. (Note that neither $\{B_i^s\}_{i\geq 1}$ nor idle $\{I_i^s\}_{i\geq 1}$ are necessarily i.i.d. random sequences.). The idle period I_i^s of the virtual channel is formed through a sequence of idle periods $\{I_{n_1}^{(c_1)}, I_{n_2}^{(c_2)}, ..., I_{n_l}^{(c_k)}\}$ from multiple channels $\{c_1, c_2, ..., c_k\}$. The actual idle time $A_i^{(j)}$ an SU can utilize from a particular idle period $I_i^{(j)}$ of channel j depends on channel switching policies, which specify whether and when the SU should switch to channel j if the current channel becomes busy. Obviously, we have $0 \leq A_i^{(j)} \leq I_i^{(j)}$ and $\{A_i^{(j)}\}_{i\geq 1}$ are independent but not necessarily equally distributed. The delay under spectrum mobility is defined as follows.

Definition 7. *(Spectrum mobility) Consider a channel* $1 \leq j \leq K$ *with busy periods* $\{B_i^{(j)}\}_{i\geq 1}$*, idle periods* $\{I_i^{(j)}\}_{i\geq 1}$*, and the corresponding actual idle times* $\{A_i^{(j)}\}_{i\geq 1}$*. During each* $A_i^{(j)}$*, we define the transmission time* $Y_i^{(j)}$ *as*

$$Y_i^{(j)} := \sup\{nL_p : nL_p \leq A_i^{(j)}\}. \tag{103}$$

Furthermore, we define

$$N_s^{(j)}(t) := \sup\left\{n_j : \sum_{i=1}^{n_j}(I_i^{(j)} + B_i^{(j)}) < t\right\}, \tag{104}$$

and the total delay $T_s(L)$ *under spectrum mobility is defined as*

$$T_s(L) := \inf\left\{t : \sum_{j=1}^{K}\sum_{i=1}^{N_s^{(j)}(t)} Y_i^{(j)} > L\right\}. \tag{105}$$

4.4.2 Multi-radio Diversity

By multi-radio diversity, an SU is equipped with K radio interfaces with each one operating on a different channel. With this feature, there exist two transmission policies: *static multi-radio diversity* and *dynamic multi-radio diversity*. Under the *static* policy, before transmitting a message, the SU divides it into K fragments with each fragment segmented into packets and sent over a preassigned interface. The total transmission delay is the time for the SU to finish sending all fragments. On the contrary, under the *dynamic* policy, without fragmenting the message before transmission, the SU directly divides the message into packets and dynamically assigns each packet to an interface whenever the channel associated with this interface is idle. The total transmission delay is the time for the SU to finish sending all the packets over multiple interfaces. The transmission delay under the two policies is defined respectively as follows.

Definition 8. *(Static multi-radio diversity) Consider a message of size* L*, which is divided into fragments of sizes* $\{r_i L\}_{K\geq i\geq 1}$ *such that* $0 \leq r_i \leq 1$ *and* $\sum_{i=1}^{K} r_i = 1$*. Let*

$T_i(r_iL)$ be the delay of sending a fragment of size r_iL over interface i. Then, the total delay $T_m^s(L)$ under static multi-radio diversity is defined as

$$T_m^s(L) := \max_{K \geq i \geq 1} T_i(r_iL) \tag{106}$$

Definition 9. *(Dynamic multi-radio diversity) Given a channel $1 \leq j \leq N$ with busy periods $\{B_i^{(j)}\}_{i \geq 1}$ and idle periods $\{I_i^{(j)}\}_{i \geq 1}$. During an idle period $I_i^{(j)}$ of the channel j, we define the transmission time $X_i^{(j)}$ as*

$$X_i^{(j)} := \sup\{nL_p : nL_p \leq I_i^{(j)}\}. \tag{107}$$

Furthermore, we define

$$N_m^{(j)}(t) := \sup\left\{ n_j : \sum_{i=1}^{n_j} (I_i^{(j)} + B_i^{(j)}) < t \right\}, \tag{108}$$

and the total delay $T_m^d(L)$ under dynamic multi-radio diversity is defined as

$$T_m^d(L) := \inf\left\{ t : \sum_{j=1}^{K} \sum_{i=1}^{N_m^{(j)}(t)} X_i^{(j)} > L \right\}. \tag{109}$$

4.4.3 Asymptotic Delay Analysis

Theorem 10. *Given K channels, where $\{B_1^{(j)}\}_{K \geq j \geq 1}$ are regularly varying random variables with indices $\alpha_1, \alpha_2, ..., \alpha_K$, respectively. Define $\alpha^{\Sigma} := \sum_{j \leq K} \alpha_j$, $\alpha^{-} := \min_{K \geq j \geq 1} \alpha_j$, and $\alpha^{+} := \max_{K \geq j \geq 1} \alpha_j$.*

1. *Under spectrum mobility, there exists a channel switching policy such that if $L \in LT$, then*

$$\lim_{t \to \infty} \frac{\log[P(T_s(L) > t)]}{\log t} \leq -\alpha^{+}. \tag{110}$$

If $L \in \mathcal{RV}(\alpha_l)$ and $E[L] < \infty$, then

$$\lim_{t \to \infty} \frac{\log[P(T_s(L) > t)]}{\log t} \leq -\min(\alpha^{+}, \alpha_l) \tag{111}$$

2. *Under static multi-radio diversity, if* $L \in LT$, *then*

$$\lim_{t\to\infty} \frac{\log[P(T_m^s(L) > t)]}{\log t} = -\alpha^- \tag{112}$$

If $L \in \mathcal{RV}(\alpha_l)$ *and* $E[L] < \infty$, *then*

$$\lim_{t\to\infty} \frac{\log[P(T_m^s(L) > t)]}{\log t} = -\min(\alpha^-, \alpha_l) \tag{113}$$

3. *Under dynamic multi-radio diversity, if* $L \in LT$,

$$\lim_{t\to\infty} \frac{\log[P(T_m^d(L) > t)]}{\log t} = -\alpha^\Sigma \tag{114}$$

If $L \in \mathcal{RV}(\alpha_l)$ *and* $E[L] < \infty$, *then*

$$\lim_{t\to\infty} \frac{\log[P(T_m^d(L) > t)]}{\log t} = -\min(\alpha^\Sigma, \alpha_l) \tag{115}$$

Remark 14. *From the above results, we see that both spectrum mobility and dynamic multi-radio diversity can greatly improve the delay performance of SUs, while static multi-radio diversity can deteriorate it. Particularly, Theorem 10(3) implies that under dynamic multi-radio diversity, the delay distribution decays at a rate equal to the sum of the indices of all channels, i.e.,* $\alpha^\Sigma := \sum_{i\le K} \alpha_i$. *This rate is much higher than the one under the single channel case, which, as implied by Theorem 6 and Corollary 1, is equal to the index* α_i *of a particular channel* i. *On the other hand, Theorem 10(1) implies that the decaying rate of the delay distribution under spectrum mobility is lower bounded by that of the best channels, which have the largest index* α_j *among all channels. On the contrary, Theorem 10(2) indicates that the delay distribution under static multi-radio diversity decays as faster as that of the worst channels, which have the smallest index* α_j *among all channels. As a consequence, compared with the single channel case, spectrum mobility and dynamic multi-radio diversity can mitigate the heavy tailed delay by increasing the orders of its finite moments at least to* $\max_{K\ge j\ge 1} \alpha_j$ *and exactly to* $\alpha^\Sigma := \sum_{i\le K} \alpha_i$, *respectively, while static multi-radio diversity can aggravate it by decreasing the orders of its finite moments to* $\min_{K\ge j\ge 1} \alpha_j$.

Corollary 2. *If $L \in \mathcal{RV}(\alpha_l)$ and $\alpha^{\Sigma} \geq \alpha_l$, then we have*

$$\lim_{t\to\infty} \frac{\log[P(T_m^d(L) > t)]}{\log t} = -\alpha_l$$

Remark 15. *This corollary directly follows from Theorem 10(3) and implies that as the number of channels increases, dynamic multi-radio diversity can achieve the optimum delay performance by maximizing the orders of finite moments. In other words, dynamic multi-radio diversity can guarantee the delay with finite moments up to order α_l, which is the highest order we can expect when transmitting heavy tailed messages of index α_l by using any multiple channel schemes .*

Corollary 2 characterizes the logarithmic asymptotics of the delay distribution for dynamic multi-radio diversity. The following Theorem 11 computes the exact asymptotic results under some confined conditions.

Theorem 11. *Given K channels, where $\{B_1^{(j)}\}_{K\geq j\geq 1}$ are regularly varying r.v.s with indices $\alpha_1, \alpha_2, ..., \alpha_K$, respectively. Define $\rho = \sum_{j=1}^{K} E[I_1^{(j)}]/(E[I_1^{(j)} + B_1^{(j)}])$ and $\alpha^* := \sum_{j\leq K:\alpha_j>1}(\alpha_j - 1)$. Assume that $L \in \mathcal{RV}(\alpha_l)$. If $\rho < 1$ and $\alpha^* > \alpha_l$, then*

$$P(T_m^d(L) > t) \sim P\left(L > \sum_{j=1}^{K} \frac{E[X_1^{(j)}]}{E[I_1^{(j)} + B_1^{(j)}]} t\right).$$

Remark 16. *The preceding result indicates that as more channels are employed, the tail distribution of the delay under dynamic multi-radio diversity is asymptotically equivalent to that of the message size L scaled by a constant.*

Proof of Theorem 10. Define $T_j(L)$ as the total delay of sending a message of size L over a particular channel $K \geq j \geq 1$. By Definition 1, we have $T_j(L) := \sum_{i=1}^{M_j}\{I_i^{(j)} + B_i^{(j)}\}$, where $M_j := \inf\{m : \sum_{i=1}^{m} X_i^{(j)} \geq L\}$ and $X_i^{(j)} := \sup\{nL_p : nL_p \leq I_i^{(j)}\}$. To prove (110) and (111), we consider a priority based channel switching policy, where if the currently used channel becomes busy, an SU always switches to the channel j^+ with the maximum index $\alpha^+ = \max_{K\geq j\geq 1} \alpha_j$ provided that this channel is idle. This

implies from (103) that $A_i^{(j)} = I_i^{(j)}$. Since the SU cannot perform channel switching in the middle of a packet being transmitted, it follows from (103) that $Z_i^{(j)} \leq Y_i^{(j)} \leq X_i^{(j)}$, where $Z_i^{(j)} := (X_i^{(j)} - L_p)\mathbf{1}(I_i^{(j)} > L_p)$, from which it follows that $T_s(L) \leq_{a.s.} T_{j^+}(L)$, where $T_{j^+}(L) := \sum_{i=1}^{M'_j}\{I_i^{(j)} + B_i^{(j)}\}$ and $M'_j := \inf\{m : \sum_{i=1}^{m} Z_i^{(j)} \geq L\}$. This implies that

$$P(T_s(L) > t) \leq P\left(T_{j^+}(L) > t\right) \tag{116}$$

If $L \in \mathrm{LT}$, using Theorem 6, we obtain $T_{j^+}(L) \in \mathcal{RV}(\alpha^+)$. This implies from (116)

$$\lim_{t\to\infty} \frac{\log[P(T_s(L) > t)]}{\log t} \leq -\alpha^+.$$

which completes the proof of (110).

If $L \in \mathcal{RV}(\alpha_l)$, from Corollary 1, we conclude

$$\lim_{t\to\infty} \frac{\log[P(T_{j^+}(L) > t)]}{\log t} = -\min(\alpha_l, \alpha^+),$$

which implies from (116) that

$$\lim_{t\to\infty} \frac{\log[P(T_s(L) > t)]}{\log t} \leq -\min(\alpha_l, \alpha^+).$$

This completes the proof of (111).

We will next prove Theorem 10(2). By the Definition of $T_m^s(L)$, we obtain

$$P(T_m^s(L) > t) = P\left(\bigcup_{i=1}^{K} T_i(r_i L) > t\right)$$

This, using the union bound, implies

$$P(T_j(r_j L) > t) \leq P(T_m^s(L) > t) \leq \sum_{j=1}^{K} P(T_j(r_j L) > t). \tag{117}$$

If L is LT, by Theorem 6, we have $T_j(r_j L) \in \mathcal{RV}(\alpha_j)$. If $\alpha_j > \alpha_i$, using Lemma 4, we obtain $P(T_j(r_j L) > t) = o(P(T_i(r_i L) > t)$. This implies from (117) that

$$-\min_{K\geq j\geq 1} \alpha_j \leq \lim_{t\to\infty} \frac{\log[P(T_m^s(L) > t)]}{\log t} \leq -\min_{K\geq j\geq 1} \alpha_j,$$

which completes the proof of (112).

If $L \in \mathcal{RV}(\alpha_l)$, from Corollary 1, we have

$$\lim_{t\to\infty} \frac{\log[P(T_j(r_j L) > t)]}{\log t} = -\min(\alpha_l, \alpha_i),$$

which implies from (117) that

$$\lim_{t\to\infty} \frac{\log[P(T_m^s(L) > t)]}{\log t} = -\min_{K\geq i\geq 1} \min(\alpha_l, \alpha_i).$$

This completes the proof of (113).

We will now prove Theorem 10(3). Let

$$T_m^{(j)}(L) := \inf\left\{ t : \sum_{i=1}^{N_m^{(j)}(t)} X_i^{(j)} > L \right\}, \tag{118}$$

which, combining (107) and (108), defines the total delay of sending a message of size L over a single channel $K \geq j \geq 1$. This implies that $T_m^{(j)}(L) \stackrel{d}{=} T_j(L)$. Since $T_m^d(L) \leq_{a.s.} T_m^{(j)}(L)\ \forall 1 \leq j \leq K$, letting $S_n^{(j)} := \sum_{i=1}^n I_i^{(j)} + B_i^{(j)}$ and $M^+ := \max_{1\leq j\leq K} M_j$, we have

$$\begin{aligned}
P(T_m^d(L) > t) &\leq P\left(\min_{1\leq j\leq K} T_m^{(j)}(L) > t\right) \\
&= P\left(\bigcap_{j=1}^{K}\left\{\sum_{i=1}^{M_j}(I_i^{(j)} + B_i^{(j)}) > t\right\}\right) \\
&\leq \sum_{n=1}^{n_0} P(M^+ = n) P\left(\bigcap_{j=1}^{K} S_n^{(j)} > t\right) \\
&\quad + \sum_{n=n_0+1}^{\infty} P(M^+ = n) P\left(\bigcap_{j=1}^{K} S_n^{(j)} > t\right) \\
&:= I + II
\end{aligned}$$

For term I, by Lemma 8(1), we have

$$\begin{aligned}
I &= \sum_{n=1}^{n_0} P(M^+ = n) \prod_{j=1}^{K} P\left(S_n^{(j)} > t\right) \\
&\sim E[(M^+)^K] \prod_{j=1}^{K} P\left(B_1^{(j)} > t\right),\ n_0 \to \infty
\end{aligned} \tag{119}$$

For term II, for any $0 < \delta < 1$, we obtain

$$\begin{aligned} II &\leq \sum_{n=n_0+1}^{\infty} P(M^+ = n) P\left(S_n^{(j)} > t\right) \\ &= \left(\sum_{n=n_0+1}^{\delta t} + \sum_{n=\delta t}^{\infty} \right) P(M^+ = n) P\left(S_n^{(j)} > t\right) \\ &:= J_1 + J_2 \end{aligned} \tag{120}$$

If $\alpha_j \leq 1$, let $\mu := 0$. Otherwise, let $\mu := E[I_1^{(j)}] + E[B_1^{(j)}]$. For $n < \delta t$, we have $y := t - n\mu > n(\delta^{-1} - \mu)$. Letting $Y_1^{(j)} := I_1^{(j)} + B_1^{(j)}$, it follows from large deviations [6] [36] theory that for any $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{y > \varepsilon n} \left| \frac{P(S_n^{(j)} - n\mu > y)}{nP(Y_1^{(j)} > y)} - 1 \right| = 0$$

This implies that there exists $C > 0$ such that as $n_0 \to \infty$

$$\limsup_{t\to\infty} J_1 \leq \lim_{n_0\to\infty} C \sum_{n=n_0+1}^{\infty} P(M^+ = n) n P(Y_1^{(j)} > y) = 0 \tag{121}$$

For term J_2, by the union bound, we have

$$J_2 \leq P(M^+ > \delta t) \leq \sum_{i=1}^{K} P(M_j > \delta t)$$

This, combining with (119), (120), and (121), proves the upper bound of $T_m^d(L)$, i.e.,

$$P(T_m^d(L) > t) \lesssim c_1 \prod_{j=1}^{K} P\left(B_1^{(j)} > t\right) + \sum_{i=1}^{K} P(M_j > \delta t)$$

where $c_1 := E[(M^+)^K]$. If $L \in LT$, by Lemma 10(3), it follows that $M_j \in LT$, which implies

$$\lim_{t\to\infty} \frac{\log[P(T_m^d(L) > t)]}{\log t} \leq -\sum_{j=1}^{K} \alpha_j. \tag{122}$$

If $L \in \mathcal{RV}(\alpha_l)$, it follows from Theorem 4 that $M_j \in \mathcal{RV}(\alpha_l)$, which implies that

$$\lim_{t\to\infty} \frac{\log[P(T_m^d(L) > t)]}{\log t} \leq -\min(\alpha_l, \sum_{j=1}^{K} \alpha_j). \tag{123}$$

As to the lower bound, by the similar arguments as the proof of upper bound, we have

$$\begin{aligned} P(T_m^d(L) > t) &\geq P\left(\bigcap_{j=1}^{K}\left\{T_j(\frac{L}{K}) > t\right\}\right) \\ &\gtrsim c_2 \prod_{j=1}^{K} P\left(B_1^{(j)} > t\right) \end{aligned} \tag{124}$$

where c_2 is a constant. Given K channels, we have $T_m^d(L) > L/K$ surely, which implies that $P(T_m^d(L) > t) > P(L/K > t)$. This, combining with (124), proves the lower bound of (114) and (115). This, in conjunction with the upper bound (122) and (123), completes the proof. □

The proof of Theorem 11 relies on Lemma 12, which corresponds to the Corollary 1.6 and Corollary 1.8 of [36].

Lemma 12. *Let $X_1, X_2, ..., X_n$ be independent random variables with $E[X_i] = 0$, for $i = 1, 2, ..., n$ and define $A_t^+ := \sum_{i=1}^n \int_{u\geq 0} u^t dP(X_i < u)$.*

1. *If $1 \geq t \geq 2$ and $A_t^+ < \infty$, then for $y^t \geq 4A_t^+$ and $x > y$*

$$P(\sum_{i=1}^{n} X_i \geq x) \leq \sum_{i=1}^{n} P(X_i > y) + \left(\frac{e^2 A_t^+}{xy^{t-1}}\right)^{x/2y}$$

2. *If $t \geq 2$ and $A_t^+ < \infty$, then*

$$P(\sum_{i=1}^{n} X_i \geq x) \leq c_t^{(1)} A_t^+ x^{-t} + exp\left\{\frac{-c_t^{(2)} x^2}{B_n^2}\right\}$$

where $c_t^{(1)} = (1 + 2/t)^t$, $c_t^{(2)} = 2(t + 2)^{-2}e^{-t}$, and $B_n^2 = \sum_{i=1}^n E[(X_i)^2]$

Proof of Theorem 11. By the definition of $T_m^d(L)$ in (109), we have

$$P(T_m^d(L) > t) = P\left(\sum_{j=1}^{K}\sum_{i=1}^{M_j(t)} X_i^{(j)} \leq L\right)$$

Define $R(t) := \sum_{j=1}^{K}\sum_{i=1}^{M_j(t)} X_i^{(j)}$. Using renewal theory, we obtain $\lim_{t\to\infty} R(t)/t = \gamma$, where $\gamma = \sum_{j=1}^{K} E[X_1^{(j)}]/(E[I_1^j + B_1^{(j)}])$. To prove Theorem 11, it is sufficient to show

that the conditions of Theorem 8 are satisfied. Since $X_1^{(j)} <_{a.s.,} I_1^{(j)}$, this implies that $E[X_1^j] < E[I_1^j]$. It follows from the assumption $\sum_{j=1}^K E[I_1^{(j)}/(E[I_1^j + B_1^{(j)}])] < 1$ that $\gamma < 1$, which verifies the first condition of Theorem 8. To verify the second condition, we define

$$\begin{cases} j^+ = \arg\max_{1\le j\le K} \left(E[I_1^{(j)} + B_1^{(j)}]\right) \\ j^- = \arg\max_{1\le j\le K} E[X_1^{(j)}] \end{cases} \tag{125}$$

Let $\varepsilon := (1-\delta)E[X_1^{(j^-)}]/\Big((1+\delta)(E[I_1^{(j^+)} + B_1^{(j^+)}])\Big)$ and $\sigma := (1-\delta)/E[I_1^{(j^+)} + B_1^{(j^+)}]$. Then, for any $0 < \delta < 1$, we obtain

$$\begin{aligned} P(R(t) < \varepsilon t) \le & P\left(\max_{1\le j\le K}\{N_m^{(j)}(t)\} \le \sigma t\right) \\ & + P\left(\sum_{j=1}^{K}\sum_{i=1}^{N_m^{(j)}(t)} X_i^{(j)} \le \varepsilon t \wedge \max_{1\le j\le N}\{N_m^{(j)}(t)\} > \sigma t\right) \\ := & J_1 + J_2 \end{aligned} \tag{126}$$

For term J_1, it follows from the independence of $\{N_m^{(j)}(t)\}_{j=1}^N$ that

$$J_1 = P(\bigcap_{j=1}^{K} N_m^{(j)}(t) \le \sigma t) = \prod_{j=1}^{K} P(N_m^{(j)}(t) \le \sigma t). \tag{127}$$

For any $B_i^{(j)}$ with $\alpha_j > 1$, let $Z_i^{(j)} := I_i^{(j)} + B_i^{(j)} - E[I_i^{(j)} + B_i^{(j)}]$. By (108), we obtain the following upper bound under the condition $\alpha_j > 1$, i.e.,

$$\begin{aligned} P(N_m^{(j)}(t) \le \sigma t) &= P\left(\sum_{i=1}^{\sigma t}(I_i^{(j)} + B_i^{(j)}) > t\right) \\ &= P\left(\sum_{i=1}^{\sigma t} Z_i^{(j)} > t - \frac{(1-\delta)(E[I_i^{(j)} + B_i^{(j)}])}{E[I_1^{(j^+)} + B_1^{(j^+)}]}t\right) \\ &\le P\left(\sum_{i=1}^{\sigma t} Z_i^{(j)} > \delta t\right) \end{aligned}$$

Since $I_1^{(j)} \in \mathrm{LT}$ and $B_1^{(j)} \in \mathcal{RV}(\alpha_j)$, an argument similar to the proof of Theorem 6 yields $P(Z_i^{(j)} > t) \sim P(B_i^{(j)} > t)$. This implies $Z_i^{(j)} \in \mathcal{RV}(\alpha_|)$. Let $\alpha_j^\Delta := \alpha_j - \Delta$. For an arbitrary small $\Delta > 0$, we have $E[(Z_i^{(j)})^{\alpha_j^\Delta}] < \infty$. If $1 \le \alpha_j \le 2$, letting $y = \delta t/2$,

an application of Lemma 12(1) and Markov inequality yields

$$P\left(\sum_{i=1}^{\sigma t} Z_i^{(j)} \geq \delta t\right) \leq \frac{2^{\alpha_j^\Delta}\sigma E[(Z_i^{(j)})^{\alpha_j^\Delta}]}{\delta^{\alpha_j^\Delta} t^{\alpha_j^\Delta - 1}} + \frac{\sigma e^2 E[(Z_i^{(j)})^{\alpha_j^\Delta}]}{\delta^{\alpha_j^\Delta} 2^{1-\alpha_j^\Delta} t^{\alpha_j^\Delta - 1}}$$
$$\leq C_j^{(1)} t^{-(\alpha_j^\Delta - 1)} \tag{128}$$

where $C_j^{(1)}$ is a constant. If $\alpha_j > 2$, by Lemma 12(2), we obtain

$$P\left(\sum_{i=1}^{\sigma t} Z_i^{(j)} > \delta t\right) \leq \frac{\sigma c_t^{(1)} E[(Z_i^{(j)})^{\alpha_j^\Delta}]}{\delta^{\alpha_j^\Delta} t^{\alpha_j^\Delta - 1}} + exp\left\{\frac{-c_t^{(2)}(\delta t)^2}{\sigma t E[(Z_i^{(j)})^2]}\right\}$$
$$\leq C_j^{(2)} t^{-(\alpha_j^\Delta - 1)} \tag{129}$$

where $C_j^{(2)}$ is a constant, $c_t^{(1)} = (1+2/\alpha_j^\Delta)^{\alpha_j^\Delta}$, and $c_t^{(2)} = 2(\alpha_j^\Delta + 2)^{-2} e^{-\alpha_j^\Delta}$. Combining (127), (128), and (129) yields

$$J_1 = o(t^{-\sum_{i \leq K: \alpha_i > 1}(\alpha_i - 1)}) = o(t^{-\alpha^*}) \tag{130}$$

For J_2, using the union bound, we obtain

$$J_2 \leq P\left(\sum_{j=1}^{K}\sum_{i=1}^{N_m^{(j)}(t)} X_i^{(j)} \leq \varepsilon t \wedge \left\{\bigcup_{l=1}^{K} N_m^{(l)}(t) > \sigma t\right\}\right)$$
$$\leq P\left(\bigcup_{l=1}^{K}\left\{\sum_{j=1}^{K}\sum_{i=1}^{N_m^{(j)}(t)} X_i^{(j)} \leq \varepsilon t \wedge N_m^{(l)}(t) > \sigma t\right\}\right)$$
$$\leq \sum_{l=1}^{K} P\left(\sum_{i=1}^{\sigma t} X_i^{(l)} \leq \varepsilon t\right)$$

Let $\widetilde{X}_i^{(l)} = E[X_i^{(l)}] - X_i^{(l)}$. By Chernoff bound, we can always find $\lambda > 0$ such that

$$P\left(\sum_{i=1}^{\sigma t} X_i^{(l)} \leq \varepsilon t\right) \leq P\left(\sum_{i=1}^{\sigma t} \widetilde{X}_i^{(l)} > \frac{\delta\sigma E[X_i^{(l)}]}{(1+\delta)} t\right)$$
$$\leq e^{-\lambda t} \tag{131}$$

which, combining with (130), implies from (126) that $P(R(t) < \varepsilon t) = o(t^{-\alpha^*})$ As a consequence, if $\alpha^+ > \alpha_l$, we obtain $P(R(t) < \varepsilon t) = o(P(L > t))$ which verities the second condition of Theorem 8 and completes the proof. □

4.5 Simulation Results

In this section, we use simulations to illustrate our theoretical results. As presented in the preceding Theorems, the SUs' HT delay is attributed to the HT message size as well as the HT PU busy time. To verify this result, we choose Pareto and exponential distributions to represent HT and LT distributions, respectively. We say that a random variable $X \in \mathcal{PAR}(\alpha, x_m)$ if X follows a Pareto distribution with parameter α and x_m, i.e., $P(X > t) = (x_m/x)^\alpha$. We say that a random variable $X \in \mathcal{EXP}(\lambda)$ if X follows an exponential distribution with parameter λ, i.e., $P(X > t) = e^{-\lambda t}$. Without loss of generality, we let packet size $L_p = 10$.

We first study the delay with both the busy time and the message size being LT. Specifically, we let $\{L, I_i\} \in \mathcal{EXP}(0.02)$ and $B_i \in \mathcal{EXP}(0.01)$. It is shown in Figure 12 that the delay tail distribution is a straight line on a y-log scale, implying that the delay is LT, specifically, exponentially distributed. Next, we investigate the cases with the HT SU message size and/or the HT PU busy periods. All the following simulation results are plotted on log-log coordinates, by which regularly varying (HT) distribution can manifest itself as a straight line.

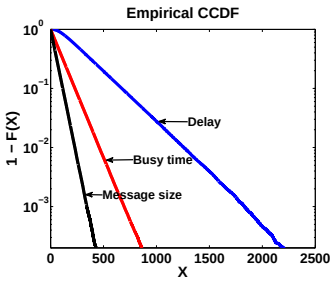

Figure 12: Delay under LT message size and LT busy time.

We next investigate the cases where either the message size or the PU busy time is HT. We first let $\{B_i, I_i\} \in \mathcal{EXP}(0.02)$ and $L \in \mathcal{PAR}(1.5, 20)$. It is seen in Figure 13 that the tail distribution of the transmission delay exhibits itself as a straight line, which is parallel to that of the message size and overlapped with the theoretical

delay tail distribution indicated by Theorem 5. This means that the transmission delay is HT and its tail distribution is as heavy as that of the message size. On the contrary, if busy time is HT while message size is LT, as indicated by Theorem 6, SUs can experience the transmission delay which has a tail distribution as heavy as that of the PU channel busy time. To verify this, we let $\{L, I_i\} \in \mathcal{EXP}(0.02)$ and $B_i \in \mathcal{PAR}(1.2, 10)$. It is seen in Figure 14 that the straight line that represents the tail distribution of the transmission delay is parallel to that of the PU busy time and coincident with the theoretical one stated by Theorem 6. This indicates that the delay tail distribution is as heavy as that of the PU busy time. In sum, Figure 13 and 14 verify Theorem 5 and 6 by showing that if one of the busy time or message size is light tailed and the other is regularly varying, then the tail of the transmission delay is asymptotically proportional to the one with regularly varying distribution.

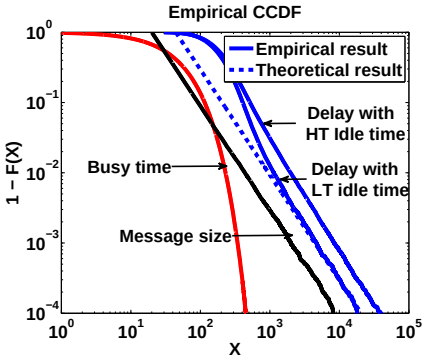

Figure 13: Delay under HT message size and LT PU busy time.

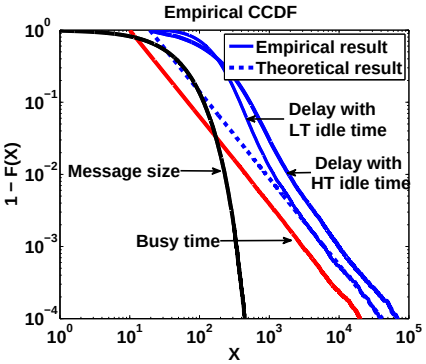

Figure 14: Delay under LT message size and HT PU busy time.

We now study the case where both message size and PU busy time are HT. In this case, Theorem 7 states that the delay performance is determined by either the busy time or the message size whichever has the heavier tail. Figure 15 shows the case where $\alpha_l > \alpha_b$ by letting $L \in \mathcal{PAR}(2, 20)$, $B_i \in \mathcal{PAR}(1.2, 10)$, and $I_i \in \mathcal{EXP}(0.02)$, while Figure 16 illustrates the case where $\alpha_l \leq \alpha_b$ by letting $L \in \mathcal{PAR}(1.2, 20)$, $B_i \in \mathcal{PAR}(2, 10)$, and $I_i \in \mathcal{EXP}(0.02)$. It is shown in Figure 15 and 16 that the tail distribution of the delay is parallel to that of either the message size or the busy time whichever has the heavier tail or smaller index, which is consistent with Theorem 7. Moreover, Figure 15 also verifies the exact asymptotic result stated in Theorem 7(1) by showing its consistence with the empirical one.

To show the impact of the HT idle time on the delay performance, we also plot the delay tail distribution with $I_i \in \mathcal{PAR}(1.2, 10)$ in Figure 13, 14, 15 and Figure 16, respectively. It can be seen that the delay tail distribution with HT idle time is parallel to the one with LT idle time in each figure. This is as expected since as indicated by Theorem 9, HT idle time has no impact on the tail heaviness of the delay.

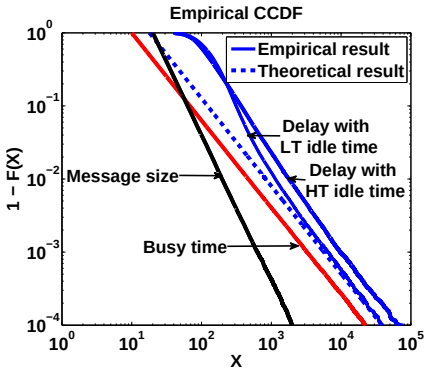

Figure 15: Delay under HT message size and HT PU busy time with $\alpha_b < \alpha_l$

We now evaluate the impact of spectrum mobility and static multi-radio diversity on the delay performance of secondary users. As indicated by Theorem 10(1) and (2), the delay under spectrum mobility is determined by the best channel which has

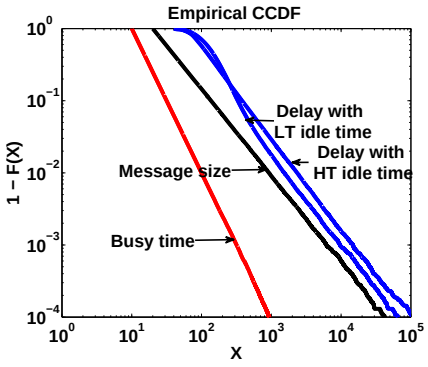

Figure 16: Delay under HT message size and HT PU busy time with $\alpha_b \geq \alpha_l$

the busy time with the lightest tail, while the delay under static multi-radio diversity is determined by the worst channel with the busy time having the heaviest tail. To verify this, we consider the scenario where there exists three PU channels with LT idle times, i.e., $\{I_i^{(1)}, I_i^{(2)}, I_i^{(3)}\} \in \mathcal{EXP}(0.01)$, and HT busy times, i.e., $B_i^{(1)} \in \mathcal{PAR}(1, 10)$, $B_i^{(2)} \in \mathcal{PAR}(0.6, 10)$, and $B_i^{(3)} \in \mathcal{PAR}(0.4, 10)$. We evaluate the delay under the case with HT message size as well as with LT message size by letting $L \in \mathcal{PAR}(2, 10)$ and $L \in \mathcal{EXP}(0.01)$, respectively. As shown in Figure 17, by taking advantage of spectrum mobility, the delay tail distribution decays faster than that of the best channel, which has the lightest tail or largest index $\alpha_1 = 1$. This implies the existence of bounded average delay. This is in sharp contrast to the delay performance of static multi-radio diversity illustrated in Figure 18, where the delay tail distribution decays as fast as the worst channel with the heaviest tail or smallest index $\alpha_3 = 0.4$. This implies that the SU will experience unbounded delay even when transmitting messages with finite mean.

We finally investigate the delay performance under dynamic multi-radio diversity. We first consider the case where there exists three PU channels with LT idle times, i.e., $\{I_i^{(1)}, I_i^{(2)}, I_i^{(3)}\} \in \mathcal{EXP}(0.01)$, and HT busy times, i.e., $B_i^{(1)} \in \mathcal{PAR}(0.4, 10)$, $B_i^{(2)} \in \mathcal{PAR}(0.5, 10)$, and $B_i^{(3)} \in \mathcal{PAR}(0.6, 10)$. Figure 19 shows the delay of sending messages with LT size, i.e., $L \in \mathcal{EXP}(0.05)$, and messages with HT size, i.e., $L \in$

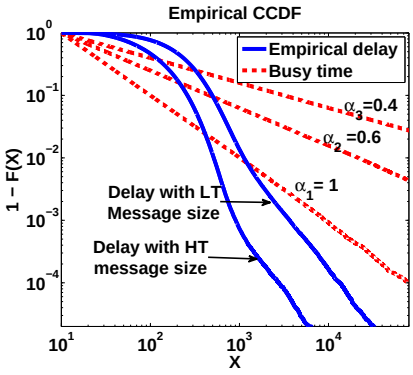

Figure 17: Delay under spectrum mobility.

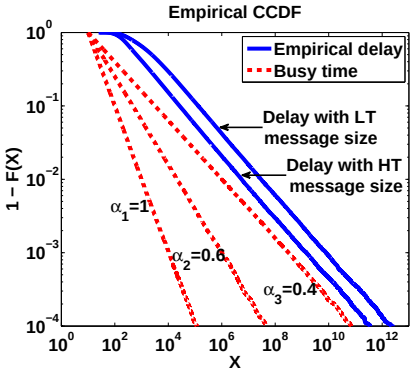

Figure 18: Delay under static multi-radio diversity.

$\mathcal{PAR}(2, 30)$, respectively. It can be seen that the delay tail distribution, as expected from Theorem 10(3), matches the baseline one which has the index of 1.5, i.e., the sum of the indices $(\alpha_1 = 0.4, \alpha_2 = 0.5, \alpha_3 = 0.6)$ of the three channels. This implies that the SU will have finite average delay, even through the average delay is unbounded if the message is transmitted on each individual channel alone. Moreover, Figure 20 shows that as the sum of indices increases and becomes larger than 2, which is the index of message size, the tail heaviness of the delay is asymptotically equivalent to that of the message size. In addition, when the sum of the indices satisfies the condition $\sum_{i=1}^{3}(\alpha_i - 1) > \alpha_l$, e.g., $\sum_{i=1}^{3} \alpha_i = 5.1$, by Theorem 11. we can obtain the exact asymptotic result of the delay tail distribution, which, as shown in Figure 20, is consistent with the empirical one.

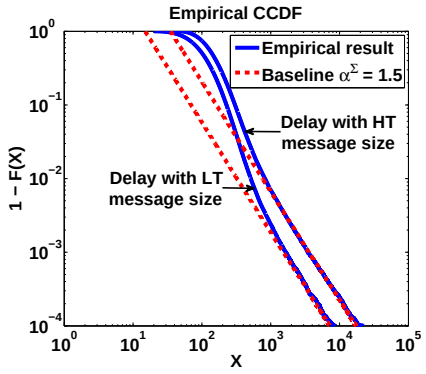

Figure 19: Delay under dynamic multi-radio diversity with $\alpha_l > \alpha^{\Sigma}$

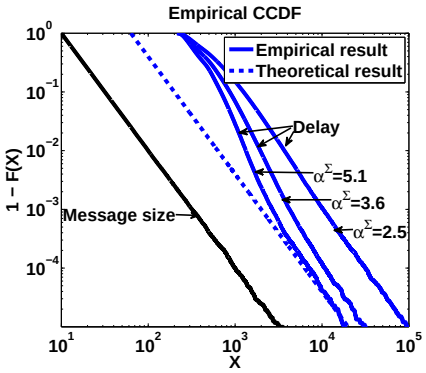

Figure 20: Delay under dynamic multi-radio diversity with $\alpha_l \leq \alpha^{\Sigma}$

CHAPTER V

THROUGHOUT-OPTIMAL SCHEDULING ALGORITHMS UNDER HEAVY-TAILED TRAFFIC

5.1 Introduction

In the previous section, we show that wireless users can experience significantly degraded delay performance under heavy-tailed environment. Based on our derived delay performance, the objective of this research is to study network stability in the presence of heavy tails and provide valuable insights for designing effective network optimization schemes. The three most common types of stability in the literature include rate stability, steady-state stability, and strong stability. [38][48]. Particularly, rate stability regulates the relationship between the time-average arrival rates and service rates, while steady-state stability demands the existence of steady-state distributions. Different from the first two definitions, strong stability has more strict criterion by requiring each queue to have finite time-average expected queue length, which is a desirable property for the applications with explicit QoS requirements.

However, under heavy-tailed environment, the conventional stability criterion may need to be revisited for facilitating the design of QoS-oriented resource allocation schemes in highly dynamic environment. Although the strong stability has been proven to be achievable in many complex systems, such stability performance is difficult to obtain in a heavy-tailed environment. Specifically, it is known that the queuing systems with heavy-tail arrival traffic or heavy-tailed service time inherently lead to heavy-tailed queueing delay, which may have unbounded expectation and variance. What is more important, recent research found that when two queues share the same server under maximum weight scheduling policy, the queue with light-tailed traffic

can experience infinite mean delay if the other queue has heavy-tailed traffic with infinite variance. Moreover, even if strong stability exists in the presence of heavy-tailed traffic, it does not necessarily imply the boundedness of the higher order moments, such as delay variance (jitter), which is one of the key metrics for the QoS-sensitive applications such as VoIP, on-line gaming and video conferencing.

The above observations motivate us to investigate network stability from a new perspective. Specifically, we introduce a new stability criterion, namely *moment stability*, which requires that all the network users with light-tailed traffic arrivals always have bounded queueing delay with finite mean and variance. Compared with strong stability, moment stability not only requires the finiteness of lower order moments, such as mean, but also demands the boundedness of higher order moments, such as variance, provided that such moments exist. What is more important, moment stability prevents heavy-tailed traffic, e.g., video conferencing and on-line gaming traffic, significantly degrading the queueing performance of light-tailed traffic, e.g., email deliveries, audio/voice traffic, and and scalar data (e.g., temperature and humidity) gathering.

Based on the definition of moment stability, the *network stability region* is defined as the closure of the set of all arrival rate vectors for which the queues of all network users can be stabilized by a feasible scheduling policy. Moreover, a scheduling policy is *throughput optimal* if it stabilizes the system for any arrival rates in the stability region. Although moment stability is a desirable property to promise QoS guaranteed applications, the conventional scheduling policies, which are effective under the light-tailed traffic, may have difficulty in achieving moment stability in the presence of heavy tails.

In this Chapter, we study the fundamental impact of heavy-tailed environment on network stability [58][56]. Towards this, we consider a dynamic spectrum access network in which multiple SUs opportunistically exploit the spectrum holes of a PU

channel. The PU channel is modeled by an alternating renewal process, which alternates between busy periods $\{B_i\}_{i\geq 1}$ and idle periods $\{I_i\}_{i\geq 1}$. Each SU is associated with an input queue and a message arrives to the queue at each time slot with a certain probability. Upon the arrival of a message with random size $L > 0$, the SU first splits it into multiple packets with constant size. At each time slot, one of the SUs can be scheduled to transmit one packet provided that the PU channel is currently detected idle. Apparently, under such generic settings, the queuing performance for the SUs has a close relationship with the message size, the PU channel availability, and the scheduling policies. For the detailed description of this model, see Section 5.2.

We first establish the critical conditions on the existence of moment stability under the exclusive access policy and the shared access policy, respectively. The former policy allows a SU has exclusive access to the PU channel without competing with other SUs, while the latter policy requires all SUs to share the PU channel. Under each policy, we study the necessary conditions for the existence of moment stability by deriving the queue length asymptotics of the SUs to determine the moment finiteness of the steady-state queue length. More specifically, it is shown that moment stability is only achievable if the heavy-tailed channel busy time has a tail index larger than three.

Utilizing this analysis, a maximum-weight-β scheduling algorithm is proposed, which associates each queue with a different parameter β and makes the scheduling decision based on the queue lengths raised to the β-th power. The maximum-weight-β scheduling can be seen as a generalized version of the celebrated maximum-weight scheduling, which makes scheduling decision based on queue lengths and is known to be throughput optimal by stabilizing the queuing system for every supportable set of traffic arrival rates [49]. However, Our asymptotic queuing analysis shows that the maximum-weight scheduling leads to the worst possible asymptotic performance for

the SU queues by letting each queue have the heaviest possible tail. In contrary, it is shown that there always exists a feasible set of β parameters such that the maximum weight-β scheduling yields the best asymptotic performance for the SU queues by letting each queue have the lightest possible tail. In this case, the maximum-weight-β schedule promises the throughput optimality by achieving moment stability for any arrival rates in the stability region.

It is worth to note that [24] and [33] are among the first research efforts to study the performance of the maximum weight-β in the queuing network and show that the maximum weight-β is effective to mitigate the impact of the queue with heavy-tailed traffic on the other queue with light-tailed traffic. Different from [24] which consider a queuing system of two users competing a single channel, we study the maximum weight-β scheduling for an arbitrary number of SUs which dynamically access a PU channel with heavy tailed behavior. In this case, SU can only access PU channel when it is detected idle and if a miss detection happens, SU needs to retransmit the collided/loss packet. Apparently, the existing literature on heavy tails do not consider such dynamic channel access schemes. What is more important, the existing work only consider the simple case with two queues and one server/channel. However, dynamic spectrum access networks generally need to support much more SUs. This greatly complicates the queueing analysis because of the correlation and dependence of the queue length of multiple SUs. Thus, the existing results can not be applied.

The rest of this section is organized as follows. In Section 5.2, we introduce system model and formally define moment stability. In Section 5.3, we present the critical conditions on the existence of moment stability. In Section 5.4, we propose maximum-weight-β scheduling and prove its throughout optimality under heavy-tailed spectrum. In Section 5.5, we explain the sufficient and necessary conditions of moment stability under light-tailed spectrum and proves the throughout optimality of maximum-weight-β scheduling.

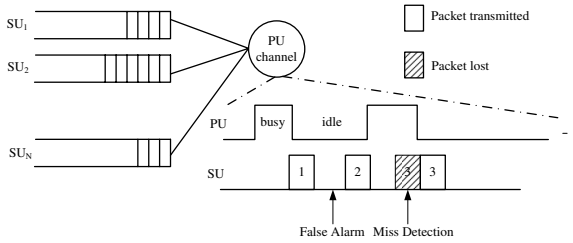

Figure 21: Multiuser dynamic spectrum access model.

5.2 Moment Stability

In this section, we first introduce the system model and then define moment stability formally. Consider N SUs sharing a PU channel, as shown in Figure 21. Time is slotted, with a unit slot length. Without loss of generality, we assume that the PU channel is of unit capacity and modeled by an alternating renewal process, which alternates between busy periods with length $\{B_i\}_{i\geq 1}$ and idle periods with length $\{I_i\}_{i\geq 1}$. $\{B_i\}_{i\geq 1}$ and $\{I_i\}_{i\geq 1}$ are mutually independent random sequences of i.i.d. random variables with distribution F_B and F_I, respectively. At each time slot, if the PU channel is detected idle, one of the SUs can be scheduled to transmit one packet per time slot. If the transmitted packet is collided with the PU transmission because of miss detection, the packet is retransmitted in the future. Assume that before the scheduling takes place, the PU channel detection result is available either through cooperative sensing or through the fusion center [2]. Let p_f denote the false alarm probability. By renewal theory, we have the service rate (throughput) of the PU channel as follows

$$\mu := \frac{(1-p_f)E[I_1]}{E[B_1]+E[I_1]}. \tag{132}$$

Let q_i denote the queue associated with SU_i. In each time slot, a message arrives to the queue q_i with a probability λ_i. Let $L_i > 0$ denote the number of packets in the message that arrives to q_i. L_i is an independent and identically distributed (i.i.d.) random variable (r.v.) from slot-to-slot, and is independent of the channel states $\{B_i\}_{i\geq 1}$ and $\{I_i\}_{i\geq 1}$. Let $A_i(t)$ denote the number of packets that arrive during time

slot t to q_i. Accordingly, the input rate Λ_i of the queue q_i is given by

$$\Lambda_i := E[A_i(t)] = \lambda_i E[L_i]. \tag{133}$$

We assume $\sum_{i=1}^{N} \Lambda_i < \mu$ so that the system is steady-state stable under any work-conserving scheduling policy, where every detected idle time slot of the PU channel is used for transmitting SUs' packets unless the SUs have empty queues. Let $Q_i(t)$ denote the queue length of q_i in time slot t. Let Q_i denote the steady-state queue length of q_i.

Definition 10. *A dynamic spectrum access network is moment stable if the steady-state queue length of every secondary user i with light-tailed arrivals has finite mean and variance, i.e., $E[Q_i] < \infty$ and $Var[Q_i] < \infty$, $\forall i \leq N, A_i(t) \in LT$.*

In the following sections, we study the existence of the throughput-optimal scheduling algorithms that can achieve moment stability by deriving the asymptotic tail distribution of the steady-state length Q_i.

5.3 Critical Conditions on the Existence of Moment Stability

In this section, we first study the existence of moment stability under the exclusive access policy and the shared access policy, respectively. The former policy allows a SU has exclusive access to the PU channel without competing with other SUs, while the latter policy requires all SUs to share the PU channel. Under each policy, we study the necessary conditions for the existence of moment stability by deriving the queue length asymptotics of the SUs to determine the moment finiteness of the steady-state queue length.

Let q_e denote the queue associated with the SU, $Q_e(t)$ the queue length at time slot t, and Q_e the steady-state queue length.

5.3.1 Main Theorems

Theorem 12. *Assume the SU message size $L \in \mathcal{RV}(\alpha_l)$ and the PU busy time $B_1 \in \mathcal{RV}(\alpha_b)$ and let $\alpha_l = \infty$ or $\alpha_b = \infty$ indicate that L or B_1 is light tailed. Then, the steady-state queue length Q_e of the SU satisfies*

$$\lim_{t\to\infty} \frac{\log[P(Q_e > t)]}{\log t} = -\min(\alpha_l, \alpha_b) + 1. \tag{134}$$

Remark 17. *The preceding results establish the relationship between the tail asymptotics of the message size L, the PU busy time B_i, and the queue length Q_e. Specifically, if either the busy time or message size is heavy tailed, then the steady-state queue length is one order heavier than the one with the heavier tail. This result implies that the expected queue length of the SUs can be infinite even if both the SU's transmitting messages and PU busy periods are of finite mean size. For example, if the message size is LT and the PU busy time is heavy tailed with tail index $2 > \alpha_b > 1$, then both the message size and the PU busy time are finite. In this case, by Theorem 12 and Lemma 1 in Section 2, the steady-state queue length of the SU has a tail index $1 > \kappa(Q_e) > 0$, which implies that both the mean and variance of Q_e are infinite, implying that moment stability is not achievable. Moreover, by Theorem 12, it is evident that the detection results of the PU channel have no impact on the asymptotic behavior of the queue length.*

By Theorem 12 and Lemma 3, the critical conditions on the existence of moniment stability is given by the following corollary.

Corollary 3. *If the channel busy time B_i has a tail index $\alpha_b < 3$, the SU necessarily has the steady-state queue length with unbounded variance, which implies the nonexistence of moment stability.*

Next, we study the critical conditions on the existence of moment stability under shared access policy by investigating the asymptotic performance of the SU queue

length under the general work conserving scheduling policies, where all the detected idle time of the PU channel are occupied for the SUs transmissions unless the SUs have empty queues.

Theorem 13. *If* $1 < \alpha_b < \min_{1\le i\le N} \alpha_{l_i}$*, then under any work conserving scheduling policy, the steady-state queue length* Q_i *of any SU* $i \le N$ *is one order heavier than the PU busy period, i.e.,*

$$\kappa(Q_i) = \alpha_b - 1, \quad \forall 1 \le i \le N \tag{135}$$

Theorem 14. *Assume* $\alpha_b \ge \min_{1\le i\le N} \alpha_{l_i} > 1$ *. Let* $\alpha^- := \min_{1\le i\le N} \alpha_{l_i}$*. Under any work-conserving policy, the steady-state queue length* Q_i *of the queue* q_i *with the smallest tail index* $\alpha_{l_i} = \alpha^-$ *follows*

$$\kappa(Q_i) = \alpha_{l_i} - 1 = \alpha^- - 1, \tag{136}$$

while the steady-state queue length Q_i *of any other queue* q_i *with* $\alpha_{l_i} > \alpha^-$ *follows*

$$\min(\alpha_{l_i}, \alpha_b) - 1 \ge \kappa(Q_i) \ge \alpha^- - 1. \tag{137}$$

Remark 18. *We can see from Theorem 13 that if the PU busy time has a heavier tail than the input traffic of any SU, then the tail asymptotic of the queue length is insensitive to the choice of a scheduling policy. In this case, under any work conserving policy, the tail distribution of the SU queue length is always one order heavier than that of the PU busy time. In contrary, by Theorem 14, if at least one of the SU queues has the input traffic with a heavier tail than the PU busy time, only the queue fed by the traffic with the heaviest tail exhibits the asymptotic behavior independent of the choice of a scheduling policy, while the other queues have bounded asymptotic performance. The above observations indicate that to guarantee the moment stability of the secondary network, the tail index* α_b *of PU busy time has to be larger than three.*

By Theorem 13 and Theorem 14, the critical conditions on the existence of moment stability under shared access policy is given by the following corollary.

Corollary 4. *If the channel busy time B_i has a tail index $\alpha_b < 3$, there exists no scheduling algorithm that can achieve moment stability, which means all SUs necessarily have the steady-state queue length with unbounded variance.*

5.3.2 Proof of the Critical Conditions for Exclusive Access Policy

To prove the critical conditions given in corollary 3, it is sufficient to prove Theorem 12.

Fictitious Queues To prove Theorem 12, we construct two fictitious queues, namely the slow queue $\widetilde{q_s}$ and the fast queue $\widetilde{q_f}$, which have the same packet arrivals, experience the same PU channel activities, and obtain the same PU channel detection results as queue q_e, but receive different services. Without loss of generality, for each queue, we assume that the first message arrives at the beginning of an idle period of the PU channel. As for the slow queue $\widetilde{q_s}$, the transmission of a new message always starts at the beginning of an idle period. This means that even if the transmission of the current message is over in the middle of an idle period, the transmission of the next message is not initiated until the next idle period begins. Thus, during the same time interval, less messages are served in the slow queue $\widetilde{q_s}$ than in the original queue q_e. This implies that

$$Q_s(t) \geq Q_e(t). \tag{138}$$

As for the fast queue $\widetilde{q_f}$, if the transmission of a message is finished in the middle of an idle period, we do not count this idle period in its service time so that each message waiting in the queue starts to be served from the beginning of the idle period during which the transmission of the previous message is finished. If a message arrives when the queue is empty, we consider two scenarios. (1) If it arrives at the beginning

of an idle period, its service time will not include the idle period during which its transmission is finished. Otherwise, (2) if it arrives in the middle of an idle period, we treat this message as if it arrives at the beginning of the idle period. It is easy to verify that during the same time interval, more messages are served in the fast queue $\widetilde{q_f}$ than in the original queue q_e. This implies that

$$Q_f(t) \leq Q_e(t). \tag{139}$$

By (138) and (139), we obtain

$$P(Q_f(t) > t) \leq P(Q_e(t) > t) \leq P(Q_s(t) > t). \tag{140}$$

We will next prove Theorem 12 by showing that the lower and upper bounds in (140) asymptotically coincide. Towards this, we derive the tail asymptotics of the steady-state queue length for the slow queue $\widetilde{q_s}$ and the fast queue $\widetilde{q_f}$, respectively.

5.3.2.1 Queue Length Asymptotics of Queue $\widetilde{q_f}$

Lemma 13. *Assume the SU message size $L \in \mathcal{RV}(\alpha_l)$ and the PU busy time $B_1 \in \mathcal{RV}(\alpha_b)$ and let $\alpha_l = \infty$ or $\alpha_b = \infty$ indicate that L or B_1 is light tailed. Then the steady-state queue length Q_f of the SU satisfies*

$$\lim_{t\to\infty} \frac{\log[P(Q_f > t)]}{\log t} = -\min(\alpha_l, \alpha_b) + 1. \tag{141}$$

To prove Lemma 13, we first define the transmission time of a message with size L in the fast queue $\widetilde{q_f}$. The construction of $\widetilde{q_f}$ indicates that the transmission attempt of a packet is always started at the beginning of an idle period. In addition, the last idle period during which the transmission is finished is excluded from the service time. Accordingly, we have the service time $T_f(L)$ of transmitting a message of size L in the queue $\widetilde{q_f}$ as follows. During an idle period I_i, let $e(j)$ denote the event that the PU channel is detected idle at time slot j and $\mathbf{1}_{e(j)}$ denote the indicator function of the event $e(j)$ where $\mathbf{1}_{e(j)} = 1$ iff the event $e(j)$ occurs.

Definition 11. *During an idle period with length I_i, the transmission time X_i of the SU is defined as*

$$X_i := \sum_{j=1}^{I_i} \mathbf{1}_{e(j)}, \tag{142}$$

the total number of idle periods the SU occupies for transmitting a message of size L,excluding the last idle period during which the transmission is finished, is defined as

$$M_f := \inf\left\{ m : \sum_{i=1}^{m} X_i \geq L \right\} - 1, \tag{143}$$

and for the fast queue $\widetilde{q_f}$, the total service (transmission) time $T_f(L)$ of a message of size L is defined as

$$T_f(L) := \sum_{i=1}^{M_f} \{I_i + B_i\}. \tag{144}$$

The tail asymptotics of the transmission time T_f is given by the following Lemma.

Lemma 14. *Assume that $B_i \in \mathcal{RV}(\alpha_b)$. If $L \in$ LT or $L \in \mathcal{RV}(\alpha_l)$ with $\alpha_b < \alpha_l$, we have*

$$P(T_f(L) > t) \sim E[M_f] P(B_1 > t). \tag{145}$$

Assume that $L \in \mathcal{RV}(\alpha_l)$ and $E[L] < \infty$.

1. *If $B_i \in$ LT, we have*

$$P(T_f(L) > t) \sim P\left(L > \frac{E[X_1]}{E[I_1] + E[B_1]} t \right). \tag{146}$$

2. *If $B_i \in \mathcal{RV}(\alpha_b)$ with $\alpha_l < \alpha_b$, we have*

$$\begin{aligned} P(T_f(L) > t) \quad \sim \quad & \left(\frac{E[B_1]}{E[X_1]} \right)^{\alpha_l} P(L > t) \\ & + \left(\frac{E[I_1]}{E[X_1]} \right)^{\alpha_l} P(L > t). \end{aligned} \tag{147}$$

3. *If $B_i \in \mathcal{RV}(\alpha_b)$ with $\alpha_l = \alpha_b$, we have*

$$\begin{aligned} P(T_f(L) > t) \quad \sim \quad & \left(\frac{E[I_1]}{E[X_1]} \right)^{\alpha_l} P(L > t) \\ & + (E[B_1])^{\alpha_l} P(L > t) \\ & + E[M_f] P(B_1 > t). \end{aligned} \tag{148}$$

Remark 19. *From the above results, we see that the tail distribution of the message transmission time is as heavy as either the SU's message size or the PU busy time, whichever has the heavier tail.*

Proof of Lemma 14. The proof follows the similar arguments of proving Theorem 4 - 7 in Chapter 3. □

We are now ready to prove Lemma 13 regarding the tail asymptotics of queue length of the fast queue $\widetilde{q_f}$.

Proof of Lemma 13. Let Q_m denote the steady-state number of messages waiting in the queue. Thus, Q_m is actually the steady-state queue length of a $GI/G/1$ queue, with the message arrival rate λ and service time $T_s(L)$. Since each message i that arrives to the queue $\widetilde{q_f}$ consists of L_i packets, the steady-state queue length Q_f satisfies

$$\sum_{i=1}^{Q_m-1} L_i \leq Q_f \leq \sum_{i=1}^{Q_m} L_i. \tag{149}$$

We next prove that the lower and upper bounds match asymptotically by considering the following three cases.

(1) If $B_i \in \mathcal{RV}(\alpha_b)$ and $L \in \text{LT}$ or $L \in \mathcal{RV}(\alpha_l)$ with $\alpha_b < \alpha_l$, it follows by Lemma 14 that the service time $T_f(L) \in \mathcal{RV}(\alpha_b)$, which implies that $T_f(L)$ is subexponentially distributed. Let $\rho = \lambda E[T_f(L)]$ is the traffic intensity. By applying Theorem 1 in [7], the steady-state waiting time W_m of a message in the queue is given by

$$P(W_m > t) \sim \frac{\rho}{1-\rho} \int_t^\infty \frac{P(T_f(L) > x)}{E[T_f(L)]} dx, \tag{150}$$

which, by distributional Little's law and regular variation, yields

$$P(Q_m > t) \sim P(\lambda W_m > t) \sim \frac{\lambda^{\alpha_b+1} E[M]}{(1-\rho)(\alpha_b - 1)} t P(B_1 > t). \tag{151}$$

This implies that $Q_m \in \mathcal{RV}(\alpha_b - 1)$. Combining (149) and (151), it follows from the sum property of random number of regularly varying random variables [17] that

$$P(Q_f > t) \sim \frac{\lambda^{\alpha_b+1} E[M] E[L]^{\alpha_b}}{(1-\rho)(\alpha_b - 1)} t P(B_1 > t). \tag{152}$$

(2) If $L \in \mathcal{RV}(\alpha_l)$ with $E[L] < \infty$ and $B_i \in \text{LT}$, by Lemma 14, we have $T_f(L) \in \mathcal{RV}(\alpha_l)$ and thus $T_f(L)$ is subexponentially distributed. By the similar arguments for the case (1), we have

$$P(Q_f > t) \sim \frac{\lambda^{\alpha_l+1}\left((E[I_1]+E[B_1])E[L]\right)^{\alpha_l}}{(1-\rho)(\alpha_b-1)E[X_1]^{\alpha_l}} tP(L>t). \tag{153}$$

(3) If $L \in \mathcal{RV}(\alpha_l)$ with $E[L] < \infty$ and $B_i \in \mathcal{RV}(\alpha_l)$ with $\alpha_l \leq \alpha_b$, this implies by Lemma 14 and the properties of slowly varying function that $T_f(L) \in \mathcal{RV}^{\alpha_l}$. By the similar arguments for the case (1) and (2), we have $T_f(L) \in \mathcal{RV}(\alpha_l - 1)$. □

By the similar techniques, it can be shown that the steady-state queue length of slow queue Q_s has the same asymptotic performance as the fast queue Q_f. This indicates by (140) that the lower and upper bounds of the queue length Q_e coincide, which completes the proof of Theorem 12.

5.3.3 Proof of the Critical Conditions for Shared Access Policy

To prove the critical conditions given in corollary 4, it is sufficient to prove Theorem 13 and 14.

Proof of Theorem 13. Since the best scheduling scheme for a particular queue q_i is to let it receive service whenever the queue is not empty. In this case, q_i behaves as if it has exclusive access to the PU channel and no other queues compete for the service. Under this scheduling policy, q_i behaves like q_e. Thus, under any work conserving policy, we have $Q_i(t) > Q_e(t)$ and thus $P(Q_i > t) > P(Q_e > t)$, which, by Theorem 12 and the assumption $\alpha_b < \min_{1\leq i\leq N} \alpha_{l_i}$, implies that the upper bound of the tail index of Q_i satisfies

$$\kappa(Q_i) \leq \min(\alpha_b, \alpha_{l_i}) - 1 \leq \alpha_b \ - 1. \tag{154}$$

Moreover, since $P(Q_i > t) \leq P(\sum_{i=1}^{N} Q_i > t)$, invoking Lemma 15, we have the lower bound of the tail index of Q_i, i.e.,

$$\kappa(Q_i) \geq \min(\min_{1\leq i\leq N} \alpha_{l_i}, \alpha_b) - 1 \geq \alpha_b \ - 1, \tag{155}$$

which agrees with the upper bound and completes the proof. □

Proof of Theorem 14. We first prove the asymptotic results in (136) regarding the queue q_i with the input process of the heaviest tail, i.e., the smallest tail index $\alpha_{l_i} = \arg\min_{1\leq i\leq N} \alpha_{l_i}$. It is evident that the queue length Q_i is stochastically dominated by the composite queue length $\sum_{i=1}^{N} Q_i$, which, by Lemma 15 and the assumption $\alpha_{l_i} \leq \alpha_b$, proves the lower bound of (136), i.e., $\kappa(Q_i) \geq \alpha_{l_i} - 1$.

As to the upper bound, we consider the best scheduling policy for q_i, which allows q_i to receive the service whenever q_i is not empty. This policy yields the best asymptotic results for the queue q_i since q_i does not have to compete with other queues for the service and thus behaves like q_e. Invoking Theorem 12, it follows from the assumption $\alpha_{l_i} \leq \alpha_b$ that the lower bound in (136) holds, i.e., $\kappa(Q_i) \leq \alpha_{l_i} - 1$, which matches the upper lower and proves (136).

Using the similar arguments, we can prove (137) by showing that the tail asymptotics of the queue length Q_i are lower bounded by those of the composite queue length and upper bounded by those of the queue q_e. The details are omitted in the interest of brevity. □

5.4 Throughput-optimal Scheduling under Heavy-tailed Spectrum

In this section, we first introduce maximum-weight-β scheduling and then prove its throughput optimality under the bursty spectrum with heavy-tailed channel busy time. To investigate the effectiveness of the scheduling policy, in this section we assume that the channel busy time has a tail index $\alpha_b > 3$, which by corollary 4, is the necessary conditions on the existence of a feasible scheduling algorithm that can achieve moment stability.

5.4.1 Maximum-Weight-β Scheduling

The maximum-weight-β scheduling works as follows. For N queues $\{q_i\}_{1\leq i\leq N}$, each queue q_i is assigned with a positive parameter β_i. During each time slot t, the queue q_i, which satisfies the condition

$$Q_i(t)^{\beta_i} = \max_{1\leq j\leq N} Q_j(t)^{\beta_j} \tag{156}$$

wins the competition and one packet from this queue is served provided that the PU channel is detected idle. Ties are broken arbitrarily. If all parameters $\{\beta_i\}_{1\leq i\leq N}$ are equivalent, maximum-weight-β scheduling becomes the conventional maximum-weight scheduling, where at each time slot, the largest queue is served.

The asymptotic analysis of the queue length distribution in this section shows that the well-known maximum-weight scheduling leads to the worst possible asymptotic behavior for the SU queues such that each queue can have a queue length with the heaviest possible tail, which indicates that as long as one SU queue has unbounded delay or variance, so do all SU queues. On the contrary, the maximum-weight-β scheduling is proven to yield the best asymptotic performance for the SU queues by letting each queue have the smallest possible tail. In other words, the maximum-weight-β schedule is asymptotically optimal because it can ensure the queue length has the same asymptotic performance as the exclusive access case, which is the best performance one can expect.

5.4.2 Throughput Optimality

Theorem 15. *Let* $\alpha_i := \min(\alpha_{l_i}, \alpha_b) > 1$, *i.e.,* $E[L_i] < \infty$ *and* $E[B] < \infty$, *and* $\alpha^- := \min_{1\leq i\leq N} \alpha_i$. *Define*

$$\alpha_i^m = \min_{1\leq j\leq N} \frac{\beta_i}{\beta_j}(\alpha_j - 1). \tag{157}$$

Under Maximum-Weight-β scheduling, the tail index of the steady-state queue length Q_i for SU i follows

$$\kappa(Q_i) = \max(\alpha_i^m, \alpha^- - 1). \tag{158}$$

Remark 20. ***(Ineffectiveness of Maximum Weight Scheduling)*** *From the above results, we see that if all parameters $\{\beta\}_{1\leq i\leq N}$ are equivalent, all queues have the same tail index as the heaviest queue which has the smallest tail index equal to $\min_{1\leq j\leq N}(\alpha_j - 1)$. This implies that the maximum-weight scheduling leads to the worst possible tail asymptotics for the SU queues so that the queue length of each queue has the lowest orders of the finite moments. In this case, if among all queues, the queue q_i is fed by the traffic with the smallest tail index $1 < \alpha_{l_i} < 2$, then under maximum-weight scheduling, all the queues have the infinite mean steady state queue length.*

Remark 21. ***(Asymptotic Optimality of Maximum Weight-β Scheduling)*** *Theorem 15 indicates that by adjusting the parameters $\{\beta_i\}_{1\leq i\leq N}$, the maximum-weight-$\beta$ scheduling can lead to the best possible asymptotic queue length performance which is as good as that under the case where the queue has the exclusive access to the PU channel. To see this, recalling Theorem 14, the best tail performance (the largest tail index) of the queue length Q_i is that $\kappa(Q_i) = \alpha_i - 1 = \min(\alpha_{l_i}, \alpha_b) - 1$. Our objective is as follows.*

$$\begin{aligned} \text{Find} \quad & \{\beta_i\}_{1\leq i\leq N} \\ \text{Such that} \quad & \alpha_i - 1 = \min_{1\leq j\leq N} \frac{\beta_i}{\beta_j}(\alpha_j - 1) \ \ \forall 1 \leq i \leq N \end{aligned}$$

One feasible solution to the above optimization problem is given by

$$\beta_i = \alpha_i - 1, \ \ \forall 1 \leq i \leq N. \tag{159}$$

The feasibility of this solution can be easily verified by inserting (159) into (157).

Corollary 5. *Assigning each queue q_i with a weight $\beta_i = \alpha_i - 1$, maximum weight-β scheduling is throughput optimal with respect to moment stability.*

Proof. This corollary holds based on Theorem 15. Specifically, we assign each q_i with LT arrivals with a weight $\beta_i = \alpha_i - 1 = \min(\alpha_{l_i}, \alpha_b) - 1$. For LT arrivals, we have $\alpha_{l_i} = \infty$. This implies $\beta_i = \alpha_b - 1$, which, combining with the assumption that the tail index of PU busy time $\alpha_b > 3$, implies that $\beta_i > 3$. It follows by Theorem 15 that the tail index $\kappa(Q_i)$ of the steady-state queue length Q_i for each q_i is equal to β_i, i.e., $\kappa(Q_i) = \beta_i > 3$. This implies that all queues with light tailed arrivals have bounded queue length with finite mean and variance, provided that the sum of total incoming rate $\Lambda_{ii\leq M}$, given in (133), is less than the channel throughout μ, given in (132). □

The proof of Theorem 15 relies on Lemma 15, 16 and 17, which we state and prove first.

Lemma 15. *Define $\alpha^- := \min_{1\leq i\leq N} \alpha_{l_i}$. We have*

$$\kappa\left(\sum_{i=1}^{N} Q_i\right) = \min(\alpha_b, \alpha^-) - 1. \tag{160}$$

Proof of Lemma 15. Consider a fictitious queue q_v which has the arrival process $A_v(t) = \sum_{i=1}^{N} A_i(t)$ and experiences the same PU channel as the original queuing system. Since $A_i(t) \in \mathcal{RV}(\alpha_{l_i})$, it implies by regular variation that the arrival process $A_v(t) \in \mathcal{RV}(\alpha^-)$, where $\alpha^- = \min_{1\leq i\leq N} \alpha_{l_i}$. Let Q_v denote the steady state queue length of q_v, It follows by Theorem 12 that $\kappa(Q_v) = \alpha^- - 1$. Let $Q_v(t)$ denote the queue length of q_v at time t. Under any work conserving policy in the original queuing system, we have $Q_v(t) = \sum_{i=1}^{N} Q_i(t)$. This implies that $Q_v = \sum_{i=1}^{N} Q_i$ and thus $\kappa(\sum_{i=1}^{N} Q_i) = \alpha^- - 1$. This completes the proof. □

Lemma 16. *Under Maximum-Weight-β scheduling, the tail index $\kappa(Q_i)$ of the steady-state queue length Q_i is lower bounded by*

$$\kappa(Q_i) \geq \alpha_i^m. \tag{161}$$

Proof of Lemma 16. For any $1 > \delta > 0$, we have

$$\begin{aligned} P(Q_i > t) &= P\left(Q_i > t \wedge \left\{\bigcap_{j\neq i} Q_j^{\frac{\beta_j}{\beta_i}} < \delta t\right\}\right) \\ &\quad + P\left(Q_i > t \wedge \left\{\bigcup_{j\neq i} Q_j^{\frac{\beta_j}{\beta_i}} \geq \delta t\right\}\right) \\ &:= I + II. \end{aligned} \tag{162}$$

As to the term I, it denotes the probability that the queue q_i has a queue length Q_i larger than t, when all the other queues have a queue length less than $(\delta t)^{\beta_i/\beta_j}$, i.e., $Q_j < (\delta t)^{\beta_i/\beta_j}$, $\forall j \neq i$. Without loss of generality, we assume that this event occurs at time 0, which means $Q_j(0)^{\beta_i/\beta_j} < \delta t$. Let $-\tau$ denote the last time when some of the queues $j \neq i$ receive service. We have two implications. (1) $Q_i(-\tau)^{\beta_i} < Q_j(-\tau)^{\beta_j}$, $\forall j \neq i$ since q_i did not receive service at time $-\tau$. (2) $Q_j(-\tau)^{\beta_j/\beta_i} < \delta t$, $\forall j \neq i$ since q_j did not receive service during the time interval $[-\tau+1, 0]$. The two implications imply that $Q_i(-\tau) < \delta t$. Thus, to ensure $Q_i(0) > t$, the number of packets accumulated in q_i during the time interval $[-\tau+1, 0]$ is at least larger than $(1-\delta)t$, i.e., $\sum_{n=-\tau+1}^{0}(A_i(n) - C_i(n)) > (1-\delta)t$, where $C_i(n)$ denotes the number of packets that depart from q_i at time n. Thus, we obtain the upper bound of I

$$\begin{aligned} I &\leq P\left(\sum_{n=-\tau+1}^{0}(A_i(n) - C_i(n)) > (1-\delta)t, \exists \tau \geq 0\right) \\ &= P\left(\sup_{\tau\geq 0} S_\tau > (1-\delta)t\right) = P\left(Q_e > (1-\delta)t\right), \end{aligned}$$

where $S_\tau := \sum_{n=-\tau+1}^{0}(A_i(n) - C_i(n))$. The last equality holds since $\sup_{\tau\geq 0} S_\tau$ is actually the event that a single server queue (a queue having exclusive access to the PU channel) has a queue length beyond $(1-\delta)t$ at time 0. It follows by Theorem 12 that

$$\kappa(Q_e) = \min(\alpha_{l_i}, \alpha_b) - 1 = \alpha_i - 1. \tag{163}$$

As to the term II, it follows by union bound that

$$\begin{aligned} II &= P\left(\bigcup_{j\neq i}\left\{Q_i > t \wedge Q_j^{\frac{\beta_j}{\beta_i}} \geq \delta t\right\}\right) \\ &\leq \sum_{j\neq i} P\left(Q_i > t \wedge Q_j^{\frac{\beta_j}{\beta_i}} \geq \delta t\right) \\ &\leq \sum_{j\neq i} P\left(Q_i + Q_j \geq (\delta t)^{\frac{\beta_i}{\beta_j}}\right). \end{aligned} \tag{164}$$

Invoking Lemma 15, we have $\kappa(Q_i + Q_j) = \min(\alpha_i, \alpha_j) - 1$. It follows from Lemma 1 that

$$\kappa((Q_i + Q_j)^{\beta_j/\beta_i}) = \min\left(\frac{\beta_i}{\beta_j}(\alpha_i - 1), \frac{\beta_i}{\beta_j}(\alpha_j - 1)\right), \tag{165}$$

which, by Lemma 1, implies that

$$\sum_{j\neq i} P\left((Q_i + Q_j)^{\frac{\beta_j}{\beta_i}} \geq (\delta t)\right) \sim P\left((Q_i + Q_{j^*})^{\frac{\beta_{j^*}}{\beta_i}} \geq (\delta t)\right) \tag{166}$$

and

$$\kappa((Q_i + Q_{j^*})^{\frac{\beta_{j^*}}{\beta_i}}) = \min_{\{1\leq j\leq N\}} \frac{\beta_i}{\beta_j}(\alpha_j - 1) = \alpha_i^m. \tag{167}$$

Combining with (162), (163), (164), and (166) yields

$$P(Q_i > t) \lesssim P\left(Q_e > (1-\delta)t\right) + P\left((Q_i + Q_{j^*})^{\frac{\beta_{j^*}}{\beta_i}} \geq (\delta t)\right) \tag{168}$$

by which, we obtain the upper bound of the steady-state queue length Q_i under two cases.

(1) If $\alpha_i - 1 < \alpha_i^m$, it follows by (163) and (167) that $\kappa(Q_e) < \kappa((Q_i + Q_{j^*})^{\frac{\beta_{j^*}}{\beta_i}})$, which, by Lemma 1, implies that

$$\sum_{j\neq i} P\left((Q_i + Q_j)^{\frac{\beta_j}{\beta_i}} \geq (\delta t)\right) = o(P\left(Q_e > (1-\delta)t\right)) \tag{169}$$

by which we obtain from (168) that

$$\limsup_{t\to\infty} \frac{[\log(P(Q_i > t))]}{\log t} \leq \limsup_{t\to\infty} \frac{\log[P\left(Q_e > (1-\delta)t\right)}{\log t}. \tag{170}$$

This implies from (163) that

$$\kappa(Q_i) \geq \kappa(Q_e) = \alpha_i - 1. \tag{171}$$

(2) If $\alpha_i - 1 \geq \alpha_i^m$, we have $\kappa(Q_e) < \kappa((Q_i + Q_{j^*})^{\frac{\beta_{j^*}}{\beta_i}})$. It follows by Lemma 1 that

$$P\left(Q_e > (1-\delta)t\right) = o\left(\sum_{j \neq i} P\left((Q_i + Q_j)^{\frac{\beta_j}{\beta_i}} \geq (\delta t)\right)\right). \tag{172}$$

This, combining (167) and (168), yields

$$\kappa(Q_i) \geq \kappa((Q_i + Q_{j^*})^{\beta_{j^*}/\beta_i}) = \alpha_i^m, \tag{173}$$

which, in conjunction with (171), completes the proof. □

Lemma 17. *Under Maximum-Weight-β scheduling, the tail index $\kappa(Q_i)$ of the steady-state queue length Q_i is upper bounded by*

$$\kappa(Q_i) \leq \alpha_i^m. \tag{174}$$

Proof of Lemma 17. To prove Lemma 17, we construct a fictitious queuing system, which consists of N queues $\{\overline{q}_i\}_{1 \leq i \leq N}$. Each queue $\overline{q}_i$ has the same input process as q_i and is associated with a dedicated PU channel i. All PU channels have the same PU activities and the same channel detection results as the PU channel in the original system. Each queue follows the same regulations as the fast queue $\widetilde{q}_f$.

Consider a particular queue $\overline{q}_i$. We let all the queues $\{\overline{q}_j\}_{j \neq i}$ except $\overline{q}_i$ have the exclusive access to their own dedicated PU channel j without competing with each other. The queue $\overline{q}_i$ receives service if and only if $\overline{Q}_i^{\beta_i} = \max_{1 \leq j \leq N} \overline{Q}_j^{\beta_j}$. In such a system, it is easy to prove that the fictitious queue $\overline{q}_i$ has shorter queue length than the queue q_i in the original system, i.e.,

$$Q_i(t) \geq \overline{Q}_i(t). \tag{175}$$

We assume that the fictitious system is in the steady state. Let p_j denote the probability that the queue $\overline{q}_{j \neq i}$ is not empty, i.e., $p_j := P(Q_j > 0)$. Let $\mathcal{E}_j$ denote the event

where $\overline{q}_j$ is not empty and all other queues are empty, i.e.,

$$\mathcal{E}_j := \left\{ \overline{Q}_j \neq 0 \wedge \bigcap_{k \neq i,j} \overline{Q}_k = 0 \right\} \tag{176}$$

and $P(\mathcal{E}_j) := p_j \prod_{k \neq i,j}(1 - p_k)$. Thus, by (175), we have the lower bound of moments of Q_i with any order d

$$E[Q_i^d] \geq \sum_{j \neq i} P(\mathcal{E}_j) E\left[\overline{Q}_i^d | \mathcal{E}_j\right]. \tag{177}$$

In the rest of the proof, we will derive the lower bound of the conditional moments $E[\overline{Q}_i^d | \mathcal{E}_j]$. We first define the following denotations for the queue $\overline{q}_j$. Assume that the event $\mathcal{E}_j$ occurs at time t. Let $L_j^r(t)$ denote the residual length of the message currently in service, which is the number of packets that belongs to this message but still remain in the queue at time t. Let $\widetilde{L}_j^r(t)$ denote the residual length of the message currently in service if the queue is served at every time slot of the PU channel. Since actually, the queue can be only served at the idle time periods of the PU channel, this implies that

$$L_j^r(t) > \widetilde{L}_j^r(t). \tag{178}$$

Let $L_j^s(t)$ denote the age of the message currently in service, which is the number of packets from this message that are already served. Let $T_j^r(t)$ and $T_j^s(t)$ denote respectively the residual and the expanded service time of the message currently in service. From renewal theory and Lemma 14, we have

$$\kappa(\widetilde{L}_j^r(t)) = \alpha_{l_j} - 1 \tag{179}$$

and if $B_1 \in \mathcal{RV}(\alpha_b)$ and $P(L_j > t) = o(P(B_1 > t))$, then

$$P(T_j^s(t) > t) \sim C_1 t P(B_1 > t), \tag{180}$$

where C_1 is a constants. Otherwise, if $L_j \in \mathcal{RV}(\alpha_{l_j})$ and $P(B_1 > t) = o(P(L_j > t))$, then

$$C_2 t P(L_j > t) \lesssim P(T_j^s(t) > t) \lesssim C_3 t P(L_j > t), \tag{181}$$

where C_2 and C_3 are some constants. By renewal theory, it follows from (180) and (181) that

$$\kappa(T_j^s(t)) = \kappa(T_j^r(t)) = \min(\alpha_{l_j}, \alpha_b) - 1 = \alpha_j - 1. \tag{182}$$

We are now ready to prove the lower bound of the conditional moments $E[\overline{Q}_i^d|\mathcal{E}_j]$. If the event $\mathcal{E}_j$ occurs, then two possible events, $\Gamma(t)$ and $\Gamma^c(t)$, occur to $\overline{Q}_i(t)$. Define $\Gamma(t) = \{\overline{Q}_i(t)^{\beta_i} \geq T_j^r(t)^{\beta_j}\}$ and its complement $\Gamma^c(t)$. If $\Gamma(t)$ occurs, we have

$$\overline{Q}_i(t) \geq T_j^r(t)^{\frac{\beta_j}{\beta_i}}. \tag{183}$$

Otherwise, if $\Gamma^c(t)$ occurs, there are two possibilities including (1) $L_j^r(t)^{\beta_j} < \overline{Q}_i(t)^{\beta_i} < T_j^r(t)^{\beta_j}$, and (2) $\overline{Q}_i(t)^{\beta_i} < L_j^r(t)^{\beta_j}$. In the case (1), it implies from (178) that

$$\overline{Q}_i(t) \geq L_j^r(t)^{\frac{\beta_j}{\beta_i}} \geq \widetilde{L}_j^r(t)^{\frac{\beta_j}{\beta_i}}. \tag{184}$$

In the case (2), let τ denote the last time before t that $\overline{q}_i$ receives service. This means that $\overline{Q}_j(\tau)^{\beta_j} < \overline{Q}_i(\tau)^{\beta_i}$. This, combining with the fact that $\overline{Q}_j(t)^{\beta_j} \geq L_j^r(t)^{\beta_j} > \overline{Q}_i(t)^{\beta_i} > \overline{Q}_i(\tau)^{\beta_i}$, implies that the burst being served at time t did not begin to receive service at τ, i.e., $t - \tau > T_j^s(t)$. This implies that

$$\overline{Q}_i(t) = \sum_{k=1}^{t-\tau} A_i(k) + \overline{Q}_i(\tau) \geq \sum_{k=1}^{T_j^s(t)} A_i(k). \tag{185}$$

Let $S_{T_j^s} := \sum_{k=1}^{T_j^s(t)} A_i(k)$. Applying Lemma 8, it follows from (180) and (181) and

$$\kappa(S_{T_j^s}) = \min(\alpha_{l_j}, \alpha_b) - 1 = \alpha_j - 1. \tag{186}$$

Let $p_\Gamma = P(\Gamma(t))$ and $p_{\Gamma^c} = P(\Gamma^c(t))$. Combining (177), (183), (184) and (185), we obtain

$$
\begin{aligned}
E[Q_i^d] &\geq \sum_{j\neq i} P(\mathcal{E}_j)\left(p_\Gamma E\left[T_j^r(t)^{\frac{d\beta_j}{\beta_i}}\right]\right. \\
&\qquad \left. + p_{\Gamma^c} E\left[\min(\widetilde{L}_j^r(t)^{\frac{d\beta_j}{\beta_i}}, (S_{T_j^s})^d)\right]\right) \\
&\geq \sum_{j\neq i} P(\mathcal{E}_j)\left(p_\Gamma E\left[T_j^r(t)^{\frac{d\beta_j}{\beta_i}}\right]\right. \\
&\qquad \left. + p_{\Gamma^c} \min(E\left[\widetilde{L}_j^r(t)^{\frac{d\beta_j}{\beta_i}}\right], E\left[(S_{T_j^s})^d\right])\right). \qquad (187)
\end{aligned}
$$

This, combining with (179), (182), and (186), implies that if the order of the moments $d \geq \min_{j\neq i} \frac{\beta_i}{\beta_j}(\alpha_j - 1)$, then at least one of the terms on the right hand of (187), is infinite, which implies

$$\kappa(Q_i) \leq \min_{j\neq i} \frac{\beta_i}{\beta_j}(\alpha_j - 1). \qquad (188)$$

Moreover, since under any working conserving scheduling policy, Q_i is lowered bounded by Q_e. This implies that

$$\kappa(Q_i) \leq \frac{\beta_i}{\beta_i}(\alpha_i - 1), \qquad (189)$$

which, combining with (188) completes the proof. □

Proof of Theorem 15. By Lemma 17 and 16, it follows that the upper and lower bounds of $\kappa(Q_i)$ matches, This, combining the fact that $\kappa(Q_i) \geq \alpha^- - 1$ by Theorem 14, completes the proof. □

5.5 Throughout-optimal Scheduling under Light-tailed Spectrum

In the previous sections, we show that maximum weight-β scheduling is throughput optimal with respect to moment stability under a single HT PU channel, in which the busy periods of PU channel follow HT distribution. In this section, we investigate the necessary and sufficient conditions under which maximum weight-β scheduling is throughput-optimal under hybrid HT and LT traffic arrivals but with LT PU channels.

We consider a dynamic spectrum access network consisting of N SUs and M PU channel, as shown in Figure 22. Time is slotted and during each time slot, only one packet is transmitted. Let $\mathbf{S} = (S_1(t), S_2(t), ..., S_M(t)$ denote the states of PU channels. $S_i(t) \in \{0, 1\}, \forall i \in 1, ..., M$ with $S_i(t) = 0$ if channel i is busy and $S_i(t) = 1$ if channel i is idle. The processes $(S_1(t), S_2(t), ..., S_M(t)$ are independent with each other and $S_i(t)$ are i.i.d. from slot to slot, distributed according to Bernoulli process with expected mean p_i, i.e., $P(S_i(t) = 1) = p_i$. At each time slot t, each secondary user $i \in \{1, 2, ..., N\}$ receives $A_i(t)$ packets, which is i.i.d. from slot to slot, and the arrival process $A_i(t), i = 1, ..., N$ is independent from each other and independent of the PU channel states. At each time slot, a scheduling/control policy allocates the detected idle channels to the secondary users with knowledge only of the current queue lengths and instantaneous channel states. Since our primary objective is to study the impact of heavy tailed traffic on network stability, we consider the scenario where the sensing errors are negligible. The above network model presents the downlink or uplink scheduling problem for the centralized networks. The practical networks represented by this model include cellular, WiFi and mesh networks with coexisting licensed and unlicensed users.

Let $Q_i(t)$ denote the number of packets in the queue q_i of secondary user i by the end of time slot i. Define $h_{ij}(t)$ as the number of packets which can be released from queue i if channel j is allocated to queue i at time slot t. Based on the above model, $h_{ij} \in \{0, 1\}, \forall i, j$. Then, the queueing dynamics of the secondary user i can be represented by

$$Q_i(t + 1) = Q_i(t) - \sum_{j=1}^{M} h_{ij}(t)S_j(t) + A_i(t) \tag{190}$$

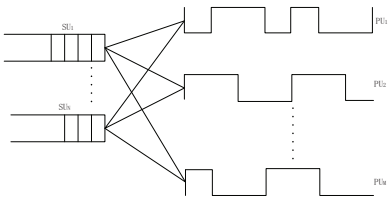

Figure 22: Multichannel dynamic spectrum access model.

subject to

$$h_{ij}(t) \in \{0,1\}, \quad \forall i,j \tag{191}$$

$$0 \leq \sum_{j=1}^{M} h_{ij}(t) \leq 1 \quad \forall i \tag{192}$$

$$0 \leq \sum_{i=1}^{N} h_{ij}(t) \leq 1 \quad \forall j \tag{193}$$

where the second constraint implies that each secondary user can only be allocated with one channel, while the third constraint means that each channel can only be assigned to one secondary user. By defining $H_i(t)$ as the number of packets which depart from queue i at time slot t under a certain control policy, the queueing dynamics in (190) can be rewritten by

$$Q_i(t+1) = Q_i(t) - H_i(t) + A_i(t) \tag{194}$$

Note that based on the above model, $H_i(t) \in \{0,1\}, \forall i = 1, ..., N$.

5.5.1 Necessary Condition of Moment Stability

In this and next sections, we derive the necessary and sufficient conditions under which there exists a feasible scheduling policy to achieve moment stability.

Theorem 16. *If there exists a scheduling policy that is throughput optimal with respect to moment stability of the system, then*

$$\sum_{i \in Q} \lambda_i \leq |Q| - \sum_{k=1}^{|Q|} P(K < k), \quad \forall Q \subset \{1, ..., N\} \tag{195}$$

where $\lambda_i = E[A_i(t)]$, $|Q|$ *denotes the cardinality of set* Q, *and* K *is the number of idle channels among total* M *channels at each time slot* t, *which follows poisson binomial distribution, denoted by* $K \sim PB(\boldsymbol{p}, M)$, $\boldsymbol{p} = (p_1, ..., p_M)$, *i.e.,*

$$P(K < k) = \sum_{l=0}^{k} \sum_{A \in F_l} \left(\prod_{j \in A} p_j \prod_{j \in A^c} (1 - p_j) \right), \tag{196}$$

where A^c *is the complementary set of* A, *and* F_l *is the collection of all subsets of* l *integers that are selectable from set* $\{1, ..., M\}$.

Remark 22. *Intuitively speaking, the right hand in* (195) *is the maximum time-average throughput the cognitive radio network can achieve, under the constraints that at each time slot, each secondary user can only access one PU channel, while each PU channel can only serve one secondary user. It can be shown that the inequality in* (195) *is also the necessary condition for the existence of the strong stability provided that such stability is achievable*

Proof. Suppose the system is moment stable under certain resource allocation policy, which, by Definition implies that the system is steady-state stable. This means for each queue, the incoming rate is equal to the service rate, i.e., $E[A_i(t)] = E[H_i(t)]$. Thus, for any subset $Q \subset \{1, .., N\}$, we have

$$\sum_{i \in Q} E[A_i(t)] = \sum_{i \in Q} E[H_i(t)] \tag{197}$$

which, by defining $K(t)$ as the number of idle channels at time slot t, can be rewritten as

$$\sum_{i \in Q} E[A_i(t)] = E\left[E\left[\sum_{i \in Q} H_i(t) | K(t), Q_i(t-1), i \in Q \right] \right] \tag{198}$$

The event $B = \{K(t), Q_i(t-1), i \in Q\}$ can be partitioned into three disjoint sets

$$\begin{aligned} B_1 &= \{K(t) = 0\} \\ B_2 &= \{K(t) = 0\}^c \wedge \{Q_i(t-1) = 0, i \in Q\} \\ B_3 &= \{K(t) = 0\}^c \wedge \{Q_i(t-1) = 0, i \in Q\}^c \end{aligned} \tag{199}$$

It is easy to verify that

$$E\left[\sum_{i\in Q} H_i(t)|B_i\right] = 0, \quad i = 1, 2 \tag{200}$$

As to event B_3, we can further divide it into two disjoint sets

$$\begin{aligned} B_3^{(1)} &= \{K(t) < |Q|\} \wedge \{Q_i(t-1) = 0, i \in Q\}^c \\ B_3^{(2)} &= \{K(t) \geq |Q|\} \wedge \{Q_i(t-1) = 0, i \in Q\}^c \end{aligned} \tag{201}$$

which, in conjunction with (200) and (198), implies

$$\sum_{i\in Q} E[A_i(t)] = E\left[\sum_{i\in Q} H_i(t)1_{B_3^{(1)}}\right] + E\left[\sum_{i\in Q} H_i(t)1_{B_3^{(2)}}\right] \tag{202}$$

Define $k_j = \{K(t) = j\} \wedge \{Q_i(t-1) = 0, i \in Q\}^c$ as the event that there are j idle channels and at least one of the queues is not empty. For the first term on the right side of (202), we have

$$\begin{aligned} E\left[\sum_{i\in Q} H_i(t)1_{B_3^{(1)}}\right] &= \sum_{j=1}^{|Q|-1} \sum_{i\in Q} E[H_i(t)|k_j]P(k_j) \\ &\leq \sum_{j=1}^{|Q|-1} jP\left(\{K(t) = j\}\right. \\ &\qquad \left. \wedge \{Q_i(t-1) = 0, i \in Q\}^c\right) \\ &\leq \sum_{j=1}^{|Q|-1} jP(K(t) = j) \end{aligned} \tag{203}$$

The second inequality is due to the fact that $\sum_{i\in Q} H_i(t) \leq K(t)$ if $K(t) \leq |Q| - 1$. For the second term on the right side of (202), we have

$$\begin{aligned} E\left[\sum_{i\in Q} H_i(t)1_{B_3^{(1)}}\right] &= \sum_{j=|Q|}^{M} \sum_{i\in Q} E[H_i(t)|k_j]P(k_j) \\ &\leq |Q| \sum_{j=|Q|}^{M} P\left(\{K(t) = j\}\right. \\ &\qquad \left. \wedge \{Q_i(t-1) = 0, i \in Q\}^c\right) \\ &\leq |Q|P(K(t) \geq |Q|) \end{aligned} \tag{204}$$

The second inequality holds because the number of channels allocated can not exceed the maximum number of queues, which implies $\sum_{i\in Q} H_i(t) \leq |Q|$ if $K(t) \geq |Q|$. Combining (202), (203), and (204), we have

$$\begin{aligned}\sum_{j\in Q}\lambda_j &\leq \sum_{j=1}^{|Q|-1} jP(K(t)=j) + |Q|P(K(t)\geq |Q|) \\ &= \sum_{j=1}^{|Q|-1}(\sum_{i=j}^{|Q|-1} P(K(t)=i) + |Q|P(K(t)\geq |Q|) \\ &= |Q| - \sum_{j=1}^{|Q|} P(K(t)<j) \end{aligned} \tag{205}$$

Since $K(t)$ follows poisson binomial distribution, this completes the proof. □

5.5.2 Sufficient Condition of Moment Stability

We next prove the sufficient condition of Moment Stability under the maximum-weight-β scheduling. More specifically, for N queues $\{q_i\}_{1\leq i\leq N}$, each queue q_i is assigned with a positive parameter β_i. During each time slot t, the scheduling algorithm chooses the channel allocation which satisfies the condition

$$\max \sum_{i,j} h_{ij}(t)Q_i(t)^{\beta_i}S_j(t) \tag{206}$$

subject to (191), (192), and (193). More specifically, we assign queues with light tailed arrivals, i.e., $A_i(t) \in LT$, with weight $\beta_i = 2$. We assign queues with heavy tailed arrivals, i.e., $A_i(t) \in \mathcal{RV}(\alpha_{l_i})$, with weight $\beta_i = \alpha_{l_i} - 1$.

Theorem 17. *The dynamic spectrum access network is moment stable, if*

$$\sum_{i\in Q}\lambda_i < |Q| - \sum_{k=1}^{|Q|} P(K<k) \quad \forall Q \subset \{1,...,N\} \tag{207}$$

where $K \sim PB(\boldsymbol{p}, M)$, $\boldsymbol{p} = (p_1, ..., p_M)$ *and each queue* i *has the* α_i*-th moment of its steady-state queue length upper bounded by*

$$E[Q_i(t)^{\beta_i}] \leq (-\frac{2N}{d})\sum_{i=1}^{N} W_i(-\frac{d}{2N}) \tag{208}$$

where $d = \max_{Q \subset \{1,...,N\}} \left\{ \sum_{i \in Q} \lambda_i - \sum_{k=1}^{|Q|} P(K > k) \right\}$ *and* $W_i()$ *follows* (213) *if* $1 \leq \alpha_i \leq \beta_i$ *and* (215) *if* $0 < \alpha_i < 1$.

Remark 23. *The Theorem above indicates that under the maximum weight-*β *scheduling algorithm, each light-tailed SU queue can achieve moment stability. The major advantage of this algorithm is that it can prevent the queues with heavy tailed arrivals from impacting the queues with light tailed arrivals. For example, assume there exist three queues, with arrival processes of* $A_1(t) \in \mathcal{RV}(1.5)$, $A_2(t) \in \mathcal{RV}(1.5)$, *and* $A_3(t) \in LT$. *By setting proper value for* β_3, *we can ensure that queue* 3 *has bounded moments of any orders even though queue* 1 *and* 2 *still have unbounded delay (since their maximum achievable order of unbounded moment is 0.5). Particularly, if letting* $2 > \beta_3 > 1$, *queue* 3 *is guaranteed to have bounded mean delay. If letting* $\beta_3 > 2$, *queue* 3 *will have bounded delay variance (jitter).*

Proof. Let $\mathcal{Q}(t) = (Q_1(t), ..., Q_N(t))$ denote a vector process of queue lengths of N secondary users. We define the Lyapunov function:

$$L(\mathcal{Q}(t)) = \sum_{i=1}^{N} L(Q_i(t)) \tag{209}$$

where

$$L(Q_i(t)) = \frac{Q_i(t)^{\beta_i+1}}{\beta_i + 1} \tag{210}$$

We next evaluate each term $L(Q_i(t))$ under two cases:$1 \leq \beta_i \leq \beta_{l_i} - 1$ and $0 < \beta_i < 1$. For the first case, using queueing dynamics in (194) and Taylor's expansions with the Lagrange form of the remainder [33], we have

$$\begin{aligned} L(Q_i(t+1)) &= \frac{1}{\beta_i + 1}(Q_i(t) + A_i(t) - H_i(t))^{\beta_i+1} \\ &= \frac{Q_i(t)^{\beta_i+1}}{\beta_i + 1} + \Delta_i(t) Q_i(t)^{\beta_i} + \beta_i \frac{\Delta_i(t)^2}{2} \delta^{\beta_i - 1} \end{aligned} \tag{211}$$

where $\Delta_i(t) = A_i(t) - H_i(t)$ and $\delta = [Q_i(t) - 1, Q_i(t) + A_i(t)]$. Therefore, by the fact that $\Delta_i(t)^2 \leq A_i(t)^2 + 1$ and $(Q_i(t) + A_i(t))^{\beta_i - 1} < 2^{\beta_i - 1}(Q_i(t)^{\beta_i - 1} + A_i(t)^{\beta_i - 1})$,

for any positive constant θ, we have

$$\begin{aligned}
&E[L_i(Q_i(t+1)) - L_i(Q(t))|\mathcal{Q}(t)] \\
&= Q_i(t)^{\beta_i} E[\Delta_i(t)|\mathcal{Q}(t)] + \frac{\beta_i}{2} E[\Delta_i(t)^2 \delta^{\beta_i - 1}|\mathcal{Q}(t)] \\
&\leq E[(A_i(t) - H_i(t))Q_i(t)^{\beta_i}|\mathcal{Q}(t)] \\
&\quad + 2^{\beta_i - 2}\beta_i E[A_i(t)^2 + 1]Q_i(t)^{\beta_i - 1} \\
&\quad + 2^{\beta_i - 2}\beta_i E[A_i(t)^{\beta_i + 1} + A_i(t)^{\beta_i - 1}] \\
&\leq E[(A_i(t) - H_i(t) + \theta)Q_i(t)^{\beta_i}|\mathcal{Q}(t)] + W_i(\theta) \qquad (212)
\end{aligned}$$

where

$$\begin{aligned}
W_i(\theta) &= \left(\theta^{-1} 2^{\beta_i - 2}\beta_i E[A_i(t)^2 + 1]\right)^{\beta_i - 1} \\
&\quad + 2^{\beta_i - 2}\beta_i E[A_i(t)^{\beta_i + 1} + A_i(t)^{\beta_i - 1}] \qquad (213)
\end{aligned}$$

The last inequality in (212) holds because $1 < \beta_i < \kappa(A_i(t)) - 1$, which implies that $E[A_i(t)^2]$, $E[A_i(t)^{\beta_i+1}]$, and $E[A_i(t)^{\beta_i-1}]$ are finite.

For the second case $0 < \beta_i < 1$, by the similar arguments, we obtain

$$\begin{aligned}
&E[L_i(Q_i(t+1)) - L_i(Q(t))|\mathcal{Q}(t)] \\
&\leq E[(A_i(t) - H_i(t) + \theta)Q_i(t)^{\beta_i}|\mathcal{Q}(t)] + W_i(\theta) \qquad (214)
\end{aligned}$$

where

$$W_i(\theta) = \theta + 1 + E[A_i(t)^{\beta_i + 1}]. \qquad (215)$$

By (209), (212), and (214), the conditional Lyapunov drift is upper bounded by

$$\begin{aligned}
&E[L(\mathcal{Q}(t+1)) - L(\mathcal{Q}(t))|\mathcal{Q}(t)] \\
&\leq \sum_{i=1}^{N} \left((\lambda_i + \theta)Q_i(t)^{\beta_i} + W_i(\theta)\right) - E\left[\sum_{i=1}^{N} H_i(t)Q_i(t)^{\beta_i}|\mathcal{Q}(t)\right] \qquad (216)
\end{aligned}$$

We now evaluate the expectation on the right side of (216). We first define the following notations. At each time slot t, we arrange the queues in a decreasing order

of their queue lengthes raised to the β_i th power, i.e., $Q_{q_1}(t)^{\beta_1}, ..., Q_{q_N}(t)^{\beta_N}$ with $Q_{q_i}(t)^{\beta_i} \geq Q_{q_{i+1}}(t)^{\beta_{i+1}}$, where ties are broken randomly. Then, we have

$$\begin{aligned}
&E\left[\sum_{i=1}^{N} H_i(t)Q_i(t)^{\beta_i}|\mathcal{Q}(t)\right] \\
&= \sum_{j=1}^{M} E\left[\sum_{i=1}^{N} H_i(t)Q_i(t)^{\beta_i}|\mathcal{Q}(t), K(t)=j\right] P(K(t)=j) \\
&= \sum_{j=1}^{M} P(K(t)=j) \sum_{i=1}^{j} Q_{q_i}(t)^{\beta_i} \\
&= \sum_{j=1}^{N} P(K(t)=j) \sum_{i=1}^{j} Q_{q_i}(t)^{\beta_i} + \sum_{j=N+1}^{M} P(K(t)=j) \sum_{i=1}^{N} Q_{q_i}(t)^{\beta_i} \\
&= \sum_{j=1}^{N} Q_{q_j}(t)^{\beta_{q_j}} \sum_{i=1}^{N} P(K(t)=i) + \sum_{j=1}^{N} Q_{q_j}(t)^{\beta_{q_j}} P(K(t)>N) \\
&= \sum_{j=1}^{N} Q_{q_j}(t)^{\beta_{q_j}} P(K(t)\geq j) \qquad (217)
\end{aligned}$$

By some computations, we can rewrite (217) as follows

$$\begin{aligned}
&\sum_{j=1}^{N} Q_{q_j}(t)^{\beta_j} P(K(t)\geq j) \\
&= \sum_{j=1}^{N-1} (Q_{q_j}(t)^{\beta_j} - Q_{q_{j+1}}(t)^{\beta_{j+1}}) \sum_{n=1}^{j} P(K(t)\geq n) \\
&+ Q_{q_N}(t)^{\beta_{q_N}} \sum_{n=1}^{N} P(K(t)\geq n) \qquad (218)
\end{aligned}$$

Similarly, we can obtain

$$\begin{aligned}
&\sum_{i=1}^{N} Q_i(t)^{\beta_i}\lambda_i = \sum_{j=1}^{N} Q_{q_j}(t)^{\beta_j}\lambda_{q_j} \\
&= \sum_{j=1}^{N-1} (Q_{q_j}(t)^{\beta_j} - Q_{q_{j+1}}(t)^{\beta_{j+1}}) \sum_{n=1}^{j} \lambda_{q_n} + Q_{q_N}(t)^{\beta_N} \sum_{n=1}^{N} \lambda_{q_n} \qquad (219)
\end{aligned}$$

Combining (218), (219), and (216), we obtain

$$\begin{aligned}
&E[L(\mathcal{Q}(t+1)) - L(\mathcal{Q}(t))|\mathcal{Q}(t)] \\
&= \sum_{j=1}^{N-1}(Q_{q_j}(t)^{\beta_j} - Q_{q_{j+1}}(t)^{\beta_{j+1}})\sum_{n=1}^{j}\lambda_{q_n} + Q_{q_N}(t)^{\beta_N}\sum_{n=1}^{N}\lambda_{q_n} \\
&\quad - \sum_{j=1}^{N-1}(Q_{q_j}(t)^{\beta_j} - Q_{q_{j+1}}(t)^{\beta_{j+1}})\sum_{n=1}^{j}P(K(t)\geq n) \\
&\quad - Q_{q_N}(t)^{\beta_{q_N}}\sum_{n=1}^{N}P(K(t)\geq n) + \sum_{i=1}^{N}\theta Q_i(t)^{\beta_i} + \sum_{i=1}^{N}W_i(\theta) \\
&\leq \sum_{j=1}^{N-1}\left((Q_{q_j}(t)^{\beta_j} - Q_{q_{j+1}}(t)^{\beta_{j+1}})\sum_{n=1}^{j}(\lambda_{q_n} - P(K(t)\geq n))\right) \\
&\quad + Q_{q_N}(t)^{\beta_N}\sum_{n=1}^{N}(\lambda_{q_n} - P(K(t)\geq n)) + \sum_{i=1}^{N}\theta Q_i(t)^{\beta_i} + \sum_{i=1}^{N}W_i(\theta) \qquad (220)
\end{aligned}$$

By defining

$$d = \max_{Q\subset\{1,\ldots,N\}}\left\{\sum_{i\in Q}\lambda_i - \sum_{k=1}^{|Q|}P(K>k)\right\} \qquad (221)$$

which is a negative constant by (207), we can rewrite (220) as

$$\begin{aligned}
&E[L(\mathcal{Q}(t+1)) - L(\mathcal{Q}(t))|\mathcal{Q}(t)] \\
&\leq dQ_{q_1}(t)^{\beta_1} + \sum_{i=1}^{N}\theta Q_i(t)^{\beta_i} + \sum_{i=1}^{N}W_i(\theta) \\
&\leq (\frac{d}{N}+\theta)\sum_{i=1}^{N}Q_i(t)^{\beta_i} + \sum_{i=1}^{N}W_i(\theta) \qquad (222)
\end{aligned}$$

The last inequality holds because q_1 has the largest β-th power queue length. Letting $\theta = -d/(2N)$, the Lyapunov drift can be bounded by

$$\begin{aligned}
&E[L(\mathcal{Q}(t+1)) - L(\mathcal{Q}(t))|\mathcal{Q}(t)] \\
&\leq (\frac{d}{2N})\sum_{i=1}^{N}Q_i(t)^{\beta_i} + \sum_{i=1}^{N}W_i(-\frac{d}{2N}) \qquad (223)
\end{aligned}$$

By Foster's criterion for ergodic Markov chain, the queueing length process converges in distribution. Using iterated mean and telescoping sums, we have

$$\sum_{i=1}^{N}E[Q_i(t)^{\beta_i}] \leq (-\frac{2N}{d})\sum_{i=1}^{N}W_i(-\frac{d}{2N}) \qquad (224)$$

where $W_i()$ is defined in (213) and (215), respectively. □

5.5.3 Throughput Optimality

The results in Theorem 16 and 17 prove the throughput optimality of maximum weight-β scheduling with respect to moment stability, which is given by the following Theorem

Theorem 18. *The sufficient and necessary conditions for moment stability is*

$$\sum_{i\in Q}\lambda_i < |Q| - \sum_{k=1}^{|Q|}(P(K<k)) \quad \forall Q \subset \{1,...,N\} \tag{225}$$

where $K \sim PB(\boldsymbol{p}, M)$, $\boldsymbol{p} = (p_1, ..., p_M)$ *and the maximum weight-*β *scheduling defined in* (206) *is throughput optimal, which stabiles any set of arrival rates within the maximum obtainable stability region given in* (225)

Remark 24. *As indicated by* (225)*, the network stability region of a cognitive radio network is characterized by the statistics of secondary user traffics, primary user activities, the number of secondary users contending the spectrum, and the total number of primary user channels available to secondary users. This region holds for any feasible work conserving policies, which utilize all idle slots of PU channels for transmissions unless secondary users have empty queues. Since work conserving policies are feasible when sensing errors are negligible,* (225) *actually provides the outer bound of the network stability region under any sensing performance.*

Remark 25. *It can be proven that the above stability region also holds for strong stability if such stability exists. In this case, the network stability regions under the two criterions overlap with each other. However, moment stability is stronger than strong stability. Specifically, if the minimum tail coefficient of all arrivals is larger than 2, both strong stability and moment stability exist, while the latter case not only guarantees the finiteness of mean but also ensures the finiteness of the higher*

order moments, such as variance. This is a important property for the QoS oriented applications such as on-line gaming and video conferencing.

CHAPTER VI

MOBILITY IMPROVES DELAY-BOUNDED CONNECTIVITY WITH HEAVY-TAILED SPECTRUM

6.1 Introduction

As a network-wide attribute, connectivity has to be maintained for reliable communications between transmitting and receiving parties in a network. For large-scale wireless networks, where full connectivity may be overly restrictive or difficult to achieve, percolation-theory based connectivity is widely adopted [15] [25][53][43]. Percolation theory concerns a phase transition phenomenon where the network exhibits fundamentally different behavior for the density λ below and above some critical density λ_c. If $\lambda > \lambda_c$, the network is percolated or in the supercritical phase and it contains a giant connected component, which consists of an infinite number of nodes. Otherwise, If $\lambda < \lambda_c$, the network is in the subcritical phase and the network is partitioned into small components containing a finite number of nodes.

Although percolation based connectivity can characterize the existence of routing paths between network devices, it does not indicate the end-to-end QoS performance, such as delay and jitter. What is more important, under heavy-tailed PU activities, the primary network can generate heavy-tailed interference region within which the secondary users will experience unbounded delay with infinite mean and/or variance [57]. Therefore, it is of significant importance to study the delay-bounded connectivity in wireless networks, which simultaneously ensures the existence of routing paths and the finiteness of transmission delay along these paths.

Mobility, as the inherent nature of today's wireless networks, has significant impact on transmission delay. Specifically, for delay-tolerant networks, mobility has

been exploited to increase network capacity at the cost of latency. Contrary to this conventional belief, we show that mobility could actually help achieve delay-bounded connectivity by improving the delay performance in wireless networks. More specifically, by exploiting the spatial diversity of the spectrum availability, mobility can allow SUs to evade from the heavy-tailed interference region induced by the primary networks. This can further guarantee the delay boundness by preventing the rise of the heavy-tailed delay. In this case, the critical question is how far secondary users need to move so that they can evade from such giant and irregular interference region as the density of the primary users increases.

This chapter aims to study the fundamental impact of heavy tailed PU activities on the delay-bounded connectivity as well as how and to what extent mobility can mitigate such impact [60][59][53]. More specifically, we show that such heavy tailed spectrum activities significantly degrade the transmission latency of secondary users. Specifically, it is proven that if the busy time of primary users is heavy tail distributed, there always exists a critical density λ_p such that if the density of primary users is larger than λ_p, the secondary users can experience unbounded average transmission latency. To encounter this, the mobility of secondary users is utilized to exploit the spatial diversity of the spectrum availability. In particular, it is shown that there exists a critical threshold on the maximum radius the secondary user can reach, above which the secondary network is percolated over time such that there exists a giant component containing an infinite number of secondary users, in which the transmission latency between any secondary users u and v is of finite mean. Moreover, it is shown that this latency scales linearly in the Euclidean distance between u and v as the distance approaches infinity.

The rest of this chapter is organized as follows. In Section 6.2, we introduce network models. In Section 6.3, we formally define delay-bounded connectivity and summarize the main results. In Section 6.4, we prove the necessary conditions on

the existence of delay-bounded connectivity in a static secondary network, where all secondary users are stationary. In Section 6.5, we derive the critical mobility radius and study the end-to-end latency in the mobile secondary networks. In Section 6.6, we present the simulation results.

6.2 Network Model

6.2.1 Heterogenous Network Architecture

We consider a heterogenous network setting, where there exist two networks: the primary network and the secondary network, where the primary network has the priority to access the spectrum. The primary users and secondary users are distributed according to the Poisson Point Process with density λ_p and λ_s, respectively, in an infinite two-dimensional space $\mathbb{R}^2$. The primary users can occupy the wireless channels whenever they have traffic to deliver, while the secondary users dynamically access the wireless channels that are not occupied by the primary users spatially or temporally. In particular, let R denote the interference range of the primary users and r denote the transmission radius of secondary users. A pair of secondary users can communicate with each other at time t if their mutual distance is less than the transmission radius r and they are outside of the interference range R of every *active* primary user at t. Specifically, the communication link between two arbitrary secondary users can be formally defined as follows.

Definition 12. *Let X_i and X_j denote the location of secondary user SU_i and SU_j, respectively. There exists a communication link between secondary user SU_i and SU_j at time t if the following conditions are fulfilled*

1. $||X_i - X_j|| < r$
2. $||X_i - Y|| > R$ *and* $||X_j - Y|| > R$, *where Y is the location of any every active primary user at time t.*

6.2.2 Primary Network Model

We model the primary network as a random disk graph denoted by $G(\lambda_p, R)$. Specifically, the primary users are distributed according to homogeneous Poisson point process with density λ_p. Let $Y_{i=1}^n$ denote the random locations of the primary users $\{1, 2, ..., n\}$. With each primary user i as the center, place a disk with radius equal to R if primary user i is active and equal to 0, otherwise.

To model the dynamic activities of primary users, we associate each primary user with an alternating renewal process, denoted by $S(t)$, which alternates between busy periods $\{B_i\}_{i\geq 1}$ and idle periods $\{I_i\}_{i\geq 1}$, which are mutually independent random sequences of i.i.d. random variables. This model has shown to be able to effectively characterize the PU behavior and thus has been widely adopted in the research field of conative radio networks [9] [18]. To effectively characterize the high burstiness in the primary user traffic, e.g., multimedia and Internet traffic, we assume the busy periods $\{B_i\}_{i\geq 1}$ follows heavy-tailed distributions, while the idle periods $\{I_i\}_{i\geq 1}$ can be either light-tail or heavy-tailed distributed.

6.2.3 Secondary Network Model

We model the secondary network as a random geometric graph denoted by $G(\lambda_s, r)$. Specifically, the secondary users are distributed according to homogeneous Poisson point process with density λ_s. Let $X_{u=1}^n$ denote the random locations of the secondary users $\{1, 2, ..., n\}$. If the mutual distance between two secondary user is less than the transmission range r, there exists a link between the two secondary users. Each link is associated with a random variable $T(L, e_{uv})$, which denotes the time of transmitting a message of random size L over link e_{uv}.

Throughout this work, we assume that the primary users are always static, while the secondary users can be mobile. Particularly, we consider that the secondary users move around in a confined region, where the maximum radius the secondary user

can reach is denoted by a. Specifically, we assume that the secondary user is moving around its initial location, which in practice can be the office or home address of the secondary users. The representative model to characterize this mobility pattern is constrained i.i.d. mobility model [26], which is formally defined as follows.

Definition 13. *(Constrained i.i.d. Mobility Model)*
Given the initial locations $\{X_1^0, ..., X_n^0\}$ *of secondary users* $u = 1, 2, ..., n$ *at time* 0. *At each time* $t = 1, 2, ...$, *the location* X_u^t *of the secondary user* u *is uniformly distributed in* $A(X_u^0, a)$, *which is the circular region centered at the initial location* X_u^0 *of* u *with radius* $a > 0$. *The positions* X_u^t *are mutually independent among all secondary users and independent of all previous locations* $X_u^{t'}, t = 0, 1, ..., t-1$.

6.3 Problem Formulation and Main Results

To study delay-bounded connectivity, we first define the transmission delay as follows. When two secondary users u and v are connected in the secondary network $G(\lambda_s, r)$ with density λ_s and node transmission range r, there exists at least one path between u and v consisting of links in the $G(\lambda_s, r)$. When u transmits a message of size L, this message can be delivered to v through different paths. For each path, we define the transmission latency between u and v as the total time the message spends traveling along this path, which is formally defined as follows.

Definition 14. *Given an arbitrary path* $l(u, v)$ *between* u *and* v, *the delay of transmitting message of size* L *over path* $l(u, v)$ *is*

$$T(u, v) = \left\{ \sum_{e_{ij} \in l(u,v)} T(L, e_{ij}) \right\}, \tag{226}$$

where $T(L, e_{ij})$ *is the delay of transmitting a message of size* L *over the link* e_{ij} *of the* $l(u, v)$.

Based on the definition of transmission latency, we define the delay-bounded connectivity for the secondary networks.

Definition 15. *A secondary network $G(\lambda_s, r)$ is connected if there exists a giant component C_∞ containing an infinite number of secondary users, and there exists at least one path between any secondary users u and v in this giant component so that the transmission latency over this path is of finite mean and variance.*

The first part of this work shows that because of the heavy tailed nature, $T(u, v)$ can have infinite mean or/and variance even if the Euclidean distance $||u-v||$ is finite.

.

Theorem 19. *Assume that the secondary network $G(\lambda_s, r)$ is percolated and the busy periods of primary users are heavy tailed with tail index α, i.e., $B_1 \in \mathcal{RV}(\alpha)$ with $\alpha < 2$. If*

$$\lambda_p > \lambda_p^* = \frac{\lambda^c}{4(R^2 - r^2/4)}, \tag{227}$$

where $1.43 < \lambda^c < 1.44$, then the delay-bounded connectivity is not achievable in the secondary network $G(\lambda_s, r)$. In other words, given any two SU u and v in the giant connected component of $G(\lambda_s, r)$, for any path between u and v, if $\alpha < 1$, then

$$E[T(u, v)] = \infty \tag{228}$$

and if $1 < \alpha < 2$, then

$$Var[T(u, v)] = \infty \tag{229}$$

provided that the Euclidean distance between u and v, $||u - v|| > d^c$, where $d^c < \infty$.

The above Theorem indicates that the heavy tailed nature of primary users can induce infinite delay mean or/and jitter between a pair of secondary users even if their mutual distance is finite. In other words, a message sent by a secondary user can be only delivered to a small portion of nodes in the secondary network within bounded delay. The intuitive explanation of this phenomenon is that since the busy periods of primary users are heavy tail distributed, the transmission time over the interfered links in $G(\lambda_s, r)$ could be heavy tailed and of infinite mean. Thus, as the

density of primary users increases, more links in $G(\lambda_s, r)$ will exhibit heavy tailed behavior so that all the pathes between two secondary users u and v will have at least one link with the transmission time of infinite mean or/and variance. This implies the infiniteness of $E[T(u,v)]$ or/and $Var[T(u,v)]$.

However, when the secondary users are mobile, there is a probability that the mean transmission latency $E[T(u,v)]$ is finite even if the busy periods of primary users are heavy tail distributed. For instance, when a secondary user u broadcasts a message, all the secondary users that reside inside the connected component of u and outside the interference range of any primary users can receive the message within finite mean delay. As time goes on, nodes move and message is passed from message-carrying nodes to new nodes whenever they are within communication range and outside the interference range of any primary users. As this process goes on, the message can be delivered within whole network without being affected by the heavy tailed activities of the primary users. The following Theorem states the sufficient condition on the maximum radius the secondary user should reach, under which the boundary d^c, defined in Theorem 19, on the maximum distance a message can travel in the secondary network vanishes.

To distinguish from the static secondary network $G(\lambda_s, r)$, we denote the mobile secondary network by $G_m(\lambda_s, r)$.

Theorem 20. *If the maximum radius a the secondary user can reach is larger than the threshold value $a^* < \infty$, where*

$$a^* = \arg\min_{a>0}((1-e^{-a^2/5\lambda_s})^2(1-(1-e^{-|R^I|\lambda_p})^{N^I}) > \frac{5}{6})$$

where $|R^I| = 2(r/\sqrt{5}+R)(2R+r/\sqrt{5})$ and $N_I = \lfloor \frac{(2/5)a^2}{R^I} \rfloor$, then the mobile secondary network $G_m(\lambda_s, r)$ achieves delay-bound connectivity such that there exists a giant component C_∞ containing an infinite number of secondary users, in which the transmission delay between any secondary users u and v with $||u-v|| < \infty$ is of finite mean

and variance, i.e., $E[T(u,v)] < \infty, \forall u, v \in C_\infty$ *and* $Var[T(u,v)] < \infty, \forall u, v \in C_\infty$

Remark 26. *By analyzing the expression* a^**, it is easy to see that* a^* *always exists under any user densities, i.e.,* λ_s *and* λ_p*, the transmission range* r *of secondary users, and the interference range* R *of primary users. This means by properly adjusting the mobility radius* a*, finite mean transmission latency can be always achieved. In other words, the maximum transmission distance* d^c *(shown in Theorem 19) induced heavy tailed spectrum activities vanishes by exploiting the mobility of secondary users.*

Now, assume the mobile secondary network $G_m(\lambda_s, r)$ achieves delay-bounded connectivity. Consider two secondary users u and v that are connected in $G_m(\lambda_s, r)$. When u broadcasts a message of size L, this message can be delivered to v through different pathes. We define the first-passage latency between u and v as the first time that v receives the message, which is formally defined as follows

Definition 16. *Consider any two secondary users* u *and* v*. The the first-passage latency from* u *to* v *is*

$$T_p(u,v) = \inf_{l(u,v)\in\mathcal{P}(u,v)} \left\{ \sum_{e_{ij}\in l(u,v)} T(L, e_{ij}) \right\} \tag{230}$$

where $l(u,v)$ *is an arbitrary path between* u *and* v *and* $\mathcal{P}(u,v)$ *is the set of paths from* u *and* v*.*

The following theorem characterizes the first-passage latency between the SU u and v.

Theorem 21. *For any secondary users* $u, v \in C_\infty$*,* $T(u,v)$ *scales linearly with the Euclidean distance between* u *and* v *as the distance approaches infinity, i.e.,*

$$\Pr\left\{ \lim_{||u-v||\to\infty} \frac{T_p(u,v)}{||u-v||} = \rho' \right\} = 1 \tag{231}$$

where ρ' *is a strictly positive value.*

6.4 Delay-bounded Connectivity in Static Wireless Networks

In this section, we prove Theorem 19 which shows that under the heavy tailed activities of primary users, the static secondary network, where all secondary users are stationary, can not achieve delay-bounded connectivity with unbounded end-to-end delay with infinite mean and variance. First, we evaluate the the lower bound of the transmission time $T(L, e_{uv})$ over a link e_{uv} in $G(\lambda_s, r)$, when the link is interfered by the transmissions of primary users. This implies that at least one of u and v is residing within the interference range R of some primary users. In this case, the link e_{uv} is available for secondary users only if all the interfering primary users are inactive. Therefore, $T(L, e_{uv})$ can be lower bounded by $T(L)$, i.e,

$$T(L, e_{uv}) > T(L), \tag{232}$$

where $T(L)$ is the time of transmitting message L over e_{uv}, provided that e_{uv} is interfered by exactly one primary user. In particular, $T(L)$ is determined by the primary user activities and can be evaluated as follows.

Suppose L is a random variable (r.v.) independent of the activities of primary users $\{B_i\}_{i\geq 1}$ and $\{I_i\}_{i\geq 1}$. For each message, the SU divides it into packets with constant size $L_p > 0$, which are then sent over the wireless channel. In each idle period I_i, the SU attempts to transmit, and if $I_i > L_p$, the SU sends packets consecutively until the remaining time of the idle period I_i is less than the packet size L_p. Otherwise, if $I_i < L_p$, the SU will wait for the next idle period for transmission. An illustration of this model is given in Figure 23. Now, the transmission time of the secondary user is formally defined as follows.

Definition 17. *During an idle period I_i, the transmission time X_i of the SU is defined as*

$$X_i := \sup\{nL_p : nL_p \leq I_i\}, \tag{233}$$

the total number of idle periods the SU occupies for transmitting a message of size L is defined as

$$M := \inf\left\{m : \sum_{i=1}^{m} X_i \geq L\right\}, \tag{234}$$

and the total delay T of the SU transmitting a message of size L is defined as

$$T(L) := \sum_{i=1}^{M-1}\{I_i + B_i\} + \left(L - \sum_{i=1}^{M-1} X_i\right) \tag{235}$$

Note that $(L - \sum_{i=1}^{M-1} X_i)$ is the exact transmission time in the last idle period.

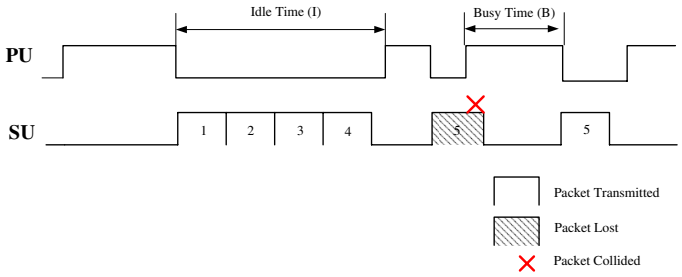

Figure 23: Dynamic spectrum access .

Assume the message size L is light-tailed distributed, we have

Lemma 18. *If $B_i \in \mathcal{RV}(\alpha_b)$ with $\alpha_b < 1$, then $E[T(L, e_{uv})] = \infty$. If $B_i \in \mathcal{RV}(\alpha_b)$ with $\alpha_b < 2$, then $Var[T(L, e_{uv})] = \infty$.*

The proof relies on the following Theorem

Theorem 22. *[57] Assume that $B_i \in \mathcal{RV}(\alpha_b)$. If $L \in$ LT, we have*

$$\lim_{t\to\infty} \frac{\log\left[P(T(L) > t)\right]}{\log t} = -\alpha_b. \tag{236}$$

Proof of Lemma 18. Since $T(L, e_{uv}) > T(L)$ surely, it follows that $P(T(L, e_{uv}) > t) > P(T(L) > t)$. By Theorem 22, we have

$$\lim_{t\to\infty} \frac{\log\left[P(T(L, e_{uv}) > t)\right]}{\log t} > -\alpha_b. \tag{237}$$

This, combining with the condition that $\alpha_b < 1$, indicates the tail coefficient of $T(L, e_{uv})$ is less than 1, implying that $E[T(L, e_{uv})] > \infty$ □

From the above results, we can see that for each link in $G(\lambda_s, r)$, if at least one of the end points of this link is within the inference range of a primary user, the mean transmission delay on this link is infinite. Based on this observation, we are ready to prove Theorem 19.

Proof of Theorem 19. Consider a secondary user u in the giant component C_u of $G(\lambda_s, r)$. To prove that $E[T(u,v)] = \infty \ \forall v \in C_u$ with $||u-v|| > d^c$, it is sufficient to prove that there exists a continuous interference region surrounding u such at all paths starting from u and ending at any v with $||u-v|| < d^c$ are disconnected by this interference region. By Lemma 18, to disconnect a path between u and v, the possible narrowest width of this interference region should be larger than the transmission radius r of the secondary user. Thus, the basic idea of the proof is to show that if condition $\lambda_p > \lambda_p^* = \frac{\lambda^c}{4(R^2 - r^2/4)}$ is satisfied, such interference region exists almost surely. Since the secondary network $G(\lambda_s, r)$ is induced by a homogeneous Poisson point process, all the nodes are probabilistically indistinguishable. We choose an arbitrary node in $G(\lambda_s, r)$ as the source node u.

We start by placing a square lattice S on $\mathbb{R}^2$, with the edge length d_p. Consider a sequence $\{G_i\}_{i \geq 1}$ of annuli around the origin. Each annulus G_i is made up of four rectangles

$$\begin{aligned} A_i^+ &= [-d_p 2^i, d_p 2^i] \times [d_p 2^{i-1}, d_p 2^i] \\ A_i^- &= [-d_p 2^i, d_p 2^i] \times [-d_p 2^i, d_p 2^{i-1}] \\ B_i^+ &= [d_p 2^{i-1}, d_p 2^i] \times [-d_p 2^i, d_p 2^i] \\ B_i^- &= [-d_p 2^i, -d_p 2^{i-1}] \times [-d_p 2^i, d_p 2^i] \end{aligned} \tag{238}$$

For each rectangle, we definite the crossing events as follows

Definition 18. *A rectangle $R = [x_1, x_2] \times [y_1, y_2]$ being crossed from left to right by a connected component in $G(\lambda_s, r)$ means that there exists a sequence of nodes*

$v_1, v_2, ..., v_m \in G(\lambda_s, r)$ *contained in* R, *with* $||X_{v_i} - X_{v_j}|| < 2\sqrt{R^2 - r^2/4}$, $i = 1, 2, ..., m-1$ *and* $0 < X_{v_1} - x_1 < 2\sqrt{R^2 - r^2/4}, 0 < x_2 - X_{v_m} < 2\sqrt{R^2 - r^2/4}$, *where* X_{v_1} *and* X_{v_m} *are the x-coordinates of nodes* v_1 *and* v_m, *respectively. A rectangle being crossed from top to bottom can be defined analogously*

We define that $A_i^+(A_i^-)$ is closed if $A_i^+(A_i^-)$ is crossed from left to right by a connected component in the primary network $G(\lambda_p, R)$, as illusrated in Figure 24. Similarly, we declare that $B_i^+(B_i^-)$ is closed if $B_i^+(B_i^-)$ is crossed from bottom to top by a connected component in $G(\lambda_p, R)$. The structure of the annulus is shown in Figure 25.

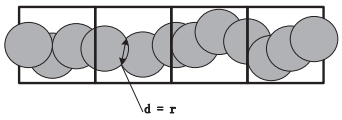

Figure 24: Closed rectangle with the narrowest width equal to the transmission radius r

By Definition 18, it is clear if A_i^+ is closed, the primary network $G(\lambda_p, R)$ generates a continuous interference region in A_i^+ with the possible narrowest width equal to the transmission radius r of the secondary users. Let $\tilde{A}_i^+$, $\tilde{A}_i^-$, $\tilde{B}_i^+$, and $\tilde{B}_i^-$ be the events that A_i^+ , A_i^- , B_i^+, and B_i^- are closed, respectively. According to RSW theorem and and the scaling property of random geometric graphs [34], when $\lambda_p > \frac{\lambda^c}{4(R^2 - r^2/4)}$, i.e., $G(\lambda_p, R)$ is in the supercritical phase, we can choose d_p large enough so that events $\tilde{A}_i^+$, $\tilde{A}_i^-$, $\tilde{B}_i^+$, and $\tilde{B}_i^-$ occur with the probability arbitrarily close to 1. This means that for any $0 < \delta < 1$, there always exists d_p' so that if $d_p \geq d_p'$, $Pr(\tilde{A}_i^+) = Pr(\tilde{A}_i^-) = Pr(\tilde{B}_i^+) = Pr(\tilde{B}_i^-) \geq \delta$

If events $\tilde{A}_i^+$, $\tilde{A}_i^-$, $\tilde{B}_i^+$, and $\tilde{B}_i^-$ occur simultaneously, the annulus G_i must contain a continuous interference region generated by the primary users, and hence all the paths starting from u are necessarily confined within the outer boundary of G_i. Denote the latter event by $\tilde{G}_i$. Since $\tilde{A}_i^+$, $\tilde{A}_i^-$, $\tilde{B}_i^+$, and $\tilde{B}_i^-$ are dependent [34], which means they

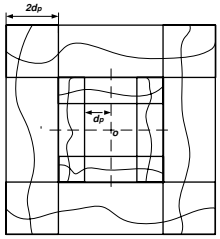

Figure 25: The annulus G_1(inside) and G_2(outside). Each annulus has four closed (crossed) rectangles

are positively correlated, utilizing Fortuin-Kasteleyn-Ginibre (FKG) inequality [34] yields

$$\begin{aligned}\Pr(\tilde{G}_i) &= \Pr(\tilde{A}_i^+ \cap \tilde{A}_i^- \cap \tilde{B}_i^+ \cap \tilde{B}_i^-) \\ &\geq \Pr(\tilde{A}_i^+)\Pr(\tilde{A}_i^-)\Pr(\tilde{B}_i^+)\Pr(\tilde{B}_i^-) \geq \delta^4\end{aligned}$$

Thus, we have $\sum_{i=1}^{\infty}\Pr(\tilde{G}_i) \geq \sum_{i=1}^{\infty}\delta^4 = \infty$. Since the construction of the annuli $\{G_i\}_{i\geq 1}$ guarantees that events $\{G_i\}_{i\geq 1}$ are independent, by the Borel-Cantelli lemma [34], there exists $j < \infty$ so that $\tilde{G}_j$ occurs with the probability 1. Since the outer boundary of annulus G_j is $[-d_p 2^j, d_p 2^j] \times [-d_p 2^j, d_p 2^j]$, this means that there exists a finite value $d^c = d_p 2^j$ such that $E[T(u,v)] = \infty \ \forall v \in C_u$ with $||u-v|| > d^c$. □

6.5 Mobility Improves Delay-bounded Connectivity

In this section, we prove Theorem 20 to show that mobility can help to achieve delay-bounded latency. To prove Theorem 20, we transform a random geometric network with mobile nodes $G_m(\lambda_s, r)$ into a random static network with stationary nodes $G(\lambda_s)$. The node positions in the latter static network are given by the initial node positions of the former mobile network. A link between a node u and v in the latter static network exists if in the former mobile network, node u and v can exchange the message within finite mean delay, without being assisted by other mobile nodes. Let $T_{uv}(L)$ define this message exchanging time between u and v without being assisted by other mobile nodes. We formally define $G(\lambda_s)$ as follows

Definition 19. *Given the initial locations* $\mathcal{X}' = \{X_1^0, X_2^0, ..., X_n^0\}$ *of secondary users* $u = 1, 2, ..., n$ *at time* 0 *and the constrained i.i.d. mobility model with radius* a*. Let* $G(\lambda_s)$ *be a static random graph, in which secondary users are located at* $\mathcal{X}'$ *and a links exists between* u *and* v *if* $E[T_{uv}(L)] < \infty$*.*

To determine the existence of a link between u and v in $G(\lambda_s)$, we first perform the following mapping processes as illustrated in Figure 26 and then prove Lemma 19. We begin by placing a square lattice $\mathcal{L}$ with the edge length D on the plane $\mathbb{R}^2$. All the vertices of L are located at $(D \times i, D \times j)$, where $(i, j) \in \mathbb{Z}$. We choose the edge length $D = a/\sqrt{5}$, where a is the maximum radius a mobile secondary user can reach. Any two squares adjacent to an edge form a *mobility rectangle* and each mobility rectangle is evenly divided into multiple *interference rectangle* with dimension $2(r/\sqrt{5}+R) \times (2R+r/\sqrt{5})$, where r is the transmission radius of secondary user and R is the interference range of primary user. At the center of each *interference rectangle*, we place a *communication rectangle* with dimension $2r/\sqrt{5} \times r/\sqrt{5}$.

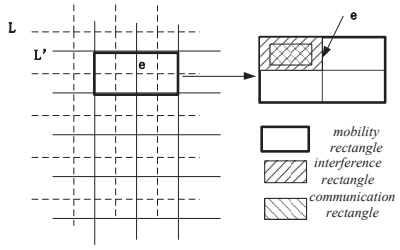

Figure 26: Lattice $\mathcal{L}'$ and its dual $\mathcal{L}$ with four interference rectangles in one mobility rectangle

Lemma 19. *Given two secondary users* u *and* v *of* $G_m(\lambda_s, r)$*. If the following two conditions are met, i.e.,*

1. *The initial location* u *and* v *at time 0 are within the same mobility rectangle.*
2. *At least one of the interference rectangles in the corresponding mobility rectangle exists no primary users.*

then there is link between u and v in $G(\lambda_S)$.

Proof. To prove the above lemma, we show that the time $T_{uv}(L)$ of exchanging a message of size L between secondary users u and v without being assisted by other mobile secondary users is of finite mean. Because the edge length of the lattice is $a/\sqrt{5}$, this ensures that if two secondary users u and v are in the mobility region at time 0, where the first condition of above Lemma is satisfied, there exists a positive probability that as time proceeds, u and v will be within the transmission range of each other. Moreover, by our lattice construction, each mobility rectangle is divided into multiple interference rectangles and each interference rectangle contains a communication rectangle. Therefore, if u and v are in the mobility rectangle at time 0, at each time slot, the probability that they can be simultaneously within a particular communication rectangle i is $p = \frac{2r^2}{5\pi R^2}$ because of the constrained i.i.d. mobility model. This means the first time $T_1(v,u)$ that u and v are within this communication rectangle follows geometric distribution with mean $E[T_1(v,u)] = p$. Assume that the interference rectangle containing communication rectangle i has no primary users, which means the second condition of the lemma is satisfied. This implies by the construction of the interference rectangle that there exists no primary user interference in the communication rectangle i. Consequently, the mean delay, $E[T_{tr}(u,,v)]$ for transmitting one packet between u and v is less than $E[T_1(v,u)] = p$, i.e., $E[T_{tr}(u,,v)] \leq E[T_1(v,u)]$. This is due to the fact that the size of the communication rectangle is $2r/\sqrt{5} \times r/\sqrt{5}$ and thus any two secondary users residing in the communication rectangle is necessarily within the transmission range of each other. Assume the mean size of the message L is finite, i.e., $E[L] < \infty$. This implies that the mean transmission delay for message L, $E[T_{uv}(L)] \leq E[L]E[T_{tr}(u,v)] \leq E[L]p$. This completes the proof. □

6.5.1 Critical Mobility Radius

Theorem 20 states that if the maximum mobility radius a is larger than the critical one a^*, then the mobile secondary network $G_m(\lambda_s, r)$ achieve delay-bounded connectivity. This means that the message broadcasted by a secondary user u can be received within the finite mean latency by an infinite number of secondary users residing at any finite distance from u. Thus, to prove Theorem 20-1 is sufficient to prove following two Lemmas

1. (Lemma 20)If the moving radius a of secondary users is larger than a threshold a^*, then $G(\lambda_s)$ is percolated

2. (Lemma 21)If $G(\lambda_s)$ is percolated, then the minimum transmission latency $T(u, v)$ between any two secondary users u and v inside the giant connected component of $G(\lambda_s)$ is of finite mean

In the rest of this subsection, we state and prove these two lemmas, respectively.

Lemma 20. *Given a secondary network $G(\lambda_s,)$, there exists a strictly positive value $a^* < \infty$, i.e.,*

$$a^* = \arg\min_{a>0}((1 - e^{-a^2/5\lambda_s})^2(1 - (1 - e^{-|R^I|\lambda_p})^{N^I}) > \frac{5}{6})$$

where $|R^I| = 2(r/\sqrt{5} + R)(2R + r/\sqrt{5})$ and $N_I = \lfloor \frac{(2/5)a^2}{R^I} \rfloor$ such that if $a > a^*$, then $G(\lambda_s)$ is percolated.

Before giving the proof of Lemma 20, we introduce several useful definitions.

Definition 20. *A vertical edge e of $\mathcal{L}$ is said to be open if the following conditions are satisfied:*

1. *At initial time $t = 0$, both squares adjacent to e contains at least one secondary user;*

2. *In the mobility region formed by the adjacent two squares, there exists at least one of the interference rectangles, which does not contain any primary users.*

Define similarly the open horizontal edge of $\mathcal{L}$ by rotating the rectangle R_e by 90 degree. Next, we construct the dual lattice of $\mathcal{L}$, which is denoted by $\mathcal{L}'$. $\mathcal{L}'$ is obtained from $\mathcal{L}$ in the following way. A vertex is placed in the center of each square of $\mathcal{L}$, and two such neighboring vertices are joined by a straight line segment. This line segment becomes an edge of $\mathcal{L}'$. As $\mathcal{L}$ is a square lattice, the dual lattice $\mathcal{L}'$ is the same lattice shifted by $(D/2, D/2)$.

Definition 21. *An edge of $\mathcal{L}'$. is said to be open if and only if its corresponding edge of $\mathcal{L}$ is open.*

Definition 22. *A path (in $\mathcal{L}$ or$\mathcal{L}'$) is said to be open if and only if all its edges are open; a path (in $\mathcal{L}$ or $\mathcal{L}'$) is said to be close if and only if all its edges are close.*

Proof of Lemma 20. The basic idea of the proof for Lemma 20 is to translate the presence of continuum percolation on $G(\lambda_s)$ into the presence of bond percolation on the lattice = $\mathcal{L}'$. More specifically, we first show that the secondary network $G(\lambda_s)$ will have an infinite connected component on the continuous plane $\mathbb{R}^2$ if bond percolation occurs on $\mathcal{L}'$, i.e.,if there exists an infinite open path on $\mathcal{L}'$.

Let E_1 and E_2 be the events when the conditions (i) and (ii) in Definition 20 are satisfied, respectively. Let C_e denote the event that an edge e is closed. The probability that C_e occurs is upper bounded by

$$\begin{aligned}\Pr(C_e) &= 1 - \Pr(E_1 \cap E_2) \stackrel{a}{=} 1 - \Pr(E_1)\Pr(E_2) \\ &= 1 - (1 - e^{-D^2\lambda_s})^2(1 - (1 - e^{-|R^I|\lambda_p})^{N^I})\end{aligned}$$

where $D = a/\sqrt{5}$, $|R^I| = 2(r/\sqrt{5} + R)(2R + r/\sqrt{5})$ is the area of the interference region R_I, and $N_I = \lfloor \frac{2D^2}{R^I} \rfloor$ is the number of interference rectangles in the mobility

region. Equality a in (239) comes from the independence of the locations of primary and secondary users.

Now, let us consider a path $P_n = \{e_i\}_{i=1}^n$ of length n in L. Because the states (i.e., open or closed) of any set of non-adjacent edges are independent, there exist at least $m \geq n/2$ edges in P_n, e.g., $\{e_j\}_{j=1}^m \subseteq \{e_i\}_{i=1}^n$, such that their states are independent from each other. Let X_{e_i} denote the event that e_i is closed. Then, the probability that the path P_n is closed is upper bounded by

$$\Pr(closed\ \mathrm{P}_n) = \Pr(\bigcap_{i=1}^{n} X_{e_i}) \leq \Pr(\bigcap_{i=1}^{m} X_{e_i}) \leq q^{\frac{n}{2}}, \tag{239}$$

where $q = Pr(C_e)$ as given in (239).

By the duality between $\mathcal{L}$ and $\mathcal{L}'$, if an open path starting from a vertex (e.g., the origin) in $\mathcal{L}'$ is finite, the origin is necessarily surrounded by a closed circuit (a closed path with the same starting and ending vertex) in the dual lattice L. Hence, by letting the latter event be O_L, the probability that there exists an infinite open path starting from the origin is $1 - Pr(O_L)$. Furthermore, from (239), we have

$$\Pr(O_{\mathcal{L}}) = \sum_{n=2}^{\infty} \sigma(n) \Pr(closed\ \mathrm{P}_{2n}) \leq \sum_{n=2}^{\infty} \sigma(n) q^n, \tag{240}$$

where $\sigma(n)$ is the number of closed circuits of the length $2n$ surrounding the origin. It is easy to show that $\sigma(n)$ is upper bounded by

$$\sigma(n) \leq (n-1)3^{2(n-1)}. \tag{241}$$

Hence, we have

$$\Pr(O_{\mathcal{L}}) = \sum_{n=2}^{\infty} (n-1)3^{2(n-1)} q^n = \frac{9q^2}{(1-9q)^2}. \tag{242}$$

Therefore, from (242) and (239), if $q = Pr(C_e) < 1/6$, or $a > a*$, where

$$a^* = \arg\min_{a>0}((1 - e^{-a^2/5\lambda_s})^2(1 - (1 - e^{-|R^I|\lambda_p})^{N^I}) > \frac{5}{6})$$

then $Pr(O_{\mathcal{L}})$ converges to a number less than one. As a consequence, the probability that there exists an infinite open path starting from the origin in $\mathcal{L}'$ is positive. According to Kolmogorovs zero-one law, this implies that an infinite path exists in $\mathcal{L}'$with probability one.

We now consider an infinite open path $P_\infty = \{e'_i\}_{i=1}^{\infty}$ in $\mathcal{L}'$. Along each edge e'_i, there exists two adjacent squares in the dual lattice $\mathcal{L}$. Therefore, along P_∞, there exists a sequence of squares $\{S_i\}_{i=1}^{\infty}$ in $\mathcal{L}$ such that any two consecutive squares, denoted by S_i and S_{i+1}, are adjacent. By Definition 20 and 21, the region comprising S_i and S_{i+1} contains at least two mutually connected SUs that belong to $G(\lambda_s)$. Thus, the sequence of squares $\{S_i\}_{i=1}^{\infty}$ forms an infinite connected component in $G(\lambda_s)$, which indicates that $G(\lambda_s)$ is percolated □

Now, we are ready to prove the first part of Theorem 20. We first introduce some useful notations. We denote by C_∞ the infinite connected component in $G(\lambda_s)$. For each coordinate $(i, 0)$ on the square lattice $\mathcal{L}$ with $i \in Z$, denote the location of the nearest secondary users in C_∞ by $\widetilde{X}_i$, i.e., $\widetilde{X}_i = \arg\min_{X_j \in C_\infty} ||X_j| - (i, 0)||$. Let $T_{m,n} = T(\widetilde{X}_m, \widetilde{X}_n)$.

Now, we assume $a > a^*$ which means $G(\lambda_s)$ is percolated. Then, to prove Theorem 20-1 is sufficient to prove following Lemma

Lemma 21. *Given two SUs, $\widetilde{X}_0$ and $\widetilde{X}_n$ within the giant connected component in $G(\lambda_s, r)$. If the mutual distance $d_{0,n} = ||\widetilde{X}_0 - \widetilde{X}_n|| < \infty$, then $E(|T_{0,n}|) = E(|T(\widetilde{X}_0, \widetilde{X}_n)|) < \infty$*

Proof. To compute the upper bound of $E(|T_{0,n}|)$, we consider the shortest path (in links) $l_{0,n}$ from $\widetilde{X}_0$ to $\widetilde{X}_n$. Denote $|l_{0,n}|$ the number of hops or links on such a path, and T_i the delay on each hop i. Since the smallest delay $T_{0,n}$ cannot be greater than the delay on any particular path, $E(|T_{0,n}|)$ is upper bounded by $E(|T_{0,n}|) \leq E(\sum_{i=1}^{|l_{0,n}|} T_i) = E(T_i)E(|l_{0,n}|)$.

By the construction of $G(\lambda_s)$ and Lemma 19, $E(T_i) < \infty$. Therefore, to show $E(|T_{0,n}|) < \infty$, is sufficient to prove $E(|l_{0,n}|) < \infty$. Let $v_{\widetilde{X}_0}$ denote the vertex of $\mathcal{L}'$, which is closest to $\widetilde{X}_0$ and Let $v_{\widetilde{X}_n}$ denote the vertex of $\mathcal{L}'$, which is closest to $\widetilde{X}_n$. By our construction of $G(\lambda_s)$, the shortest path $l_{0,n}$ between $\widetilde{X}_0$ and $\widetilde{X}_n$ is the open path on $\mathcal{L}'$ consisting of the smallest number of edges. To evaluate $|l_{0,n}|$, we construct a new square lattice on the top of square lattice $\mathcal{L}'$ (shown in fig 26), This new square lattice has the edge length $d_{0,n} = nL_D$, where L_D the edge length of $\mathcal{L}'$. Then, a sequence annuli $\{G_i(d_{0,n})\}_{i\geq 1}$ around the origin is constructed in the same way as illustrated in Figure 25. More specifically, each annulus $G_i(d_{0,n})$ is made up of four rectangles $A_i^+(d_{0,n})$,$A_i^-(d_{0,n})$, $B_i^+(d_{0,n})$, and $B_i^-(d_{0,n})$, according to equation (238), in which we substitute $d_{0,n}/2$ for d_p. We define the crossing events for the rectangles as follows.

Definition 23. *A rectangle $[x_1, x_2] \times [y_1, y_2]$ is crossed from left to right if $[x_1, x_2] \times [y_1, y_2]$ contains an open path on the square lattice $\mathcal{L}'$ that joins the left and right borders of $[x_1, x_2] \times [y_1, y_2]$. A rectangle $[x_1, x_2] \times [y_1, y_2]$ being crossed from top to bottom can be defined analogously.*

Let $\widetilde{A}_i^+(d_{0,n})$,$\widetilde{A}_i^-(d_{0,n})$, $\widetilde{B}_i^+(d_{0,n})$, and $\widetilde{B}_i^-(d_{0,n})$ be the event that $A_i^+(d_{0,n})$,$A_i^-(d_{0,n})$, $B_i^+(d_{0,n})$, and $B_i^-(d_{0,n})$ are open, respectively. By RSW theorem for independent bound percolation [20], if the states of edges of $\mathcal{L}'$ are independent and the open probability of the edge is larger than 1/2, for any $0 < \delta < 1$, there always exists $i < \infty$ such that $P(\widetilde{A}_i^+(d_{0,n})) > \delta$. Although the adjacent edges of $\mathcal{L}'$ are dependent, by the definition 20 and 21, the open states of adjacent edges in $\mathcal{L}'$ are increasing events. Moreover, by the proof of Lemma 20, the open probability of the edge is larger than 5/6, provided that $a > a^*$. This implies that if $a > a^*$, $P(\widetilde{A}_i^+(d_{0,n})) > \delta$ still holds. Denote $\widetilde{G}_i(d_{0,n})$ the event that all the four rectangles of an annulus $G_i(d_{0,n})$ are open. Since $\widetilde{A}_i^+(d_{0,n})$,$\widetilde{A}_i^-(d_{0,n})$, $\widetilde{B}_i^+(d_{0,n})$ and $\widetilde{B}_i^-(d_{0,n})$ are increasing events, by the FKG inequality [34], we have $Pr(\widetilde{G}_i(d_{0,n})) = Pr(\widetilde{A}_i^+(d_{0,n}) \cap \widetilde{A}_i^-(d_{0,n}) \cap \widetilde{B}_i^+(d_{0,n}) \cap$

$\widetilde{B}_i^-(d_{0,n})) \geq \delta^4$

If all the four rectangles of an annulus $G_i(d_{0,n})$ are open, i.e., $\widetilde{G}_i(d_{0,n})$ occurs, then the annulus $G_i(d_{0,n})$ contains an open circuit. Consequentially, the shortest path $L_{0,n}$ is necessarily included in the square $[-d_{0,n}2^i, d_{0,n}2^i] \times [-d_{0,n}2^i, d_{0,n}2^i]$. Thus, the number of hops on $l0, n$, e.g., $|l_{0,n}|$, can be upper bounded by a certain value. Since $l_{0,n}$ is included in a square of the size of $4^i d_{0,n}^2$, $|l_{0,n}|$ is upper bounded by $4^i n^2 + 2^i n$. Therefore, we have

$$Pr(|l_{0,n}| > 4^i n^2 + 2^i n) < Pr(\bigcap_{j=1}^{i} \widetilde{G}_j(d_{0,n})) < (1-\delta^4)^i$$

$E(|L_{0,n}|)$ can be upper bounded by

$$\begin{aligned} E(|l_{0,n}|) &= \sum_{k=0}^{N-1} Pr(|l_{0,n}|) > k + \sum_{k=N}^{\infty} Pr(|l_{0,n}|) > k \\ &\leq 4^{N-1}n^2 + 2^{N-1}n + \sum_{k=N}^{\infty} \frac{4^k n^2 + 2^k n}{(1-\delta^4)^{-k}} < \infty \end{aligned}$$

where N is the minimum value of the annulus index i such that $Pr(\widetilde{A}_i^+(d_{0,n}) > \delta > (\frac{3}{4})^{1/4}$. This completes the proof. □

6.5.2 First-passage Latency

Next, we prove the Theorem 21 regarding the linear scaling property of the first-passage latency $T_p(u, v)$ with the Euclidean distance between u and v as the distance approaches ∞.

The proof of Theorem 20-2 relies on the following lemma.

Lemma 22.

$$\lim_{n\to\infty} (\frac{T_{0,n}}{n}) = \rho \tag{243}$$

with probability one

where $\rho = \lim_{n\to\infty} \frac{E(T_{0,n})}{n} = \inf_{n\geq 1} \frac{E(T_{0,n})}{n}$

The main tools to prove Lemma 22 is Liggetts subadditive ergodic theorem [30], which states as follows.

Theorem 23. *(subadditive ergodic theorem) Let* $\{T_{m,n}\}$ *be a collection of random variables indexed by integers satisfying* $0 \leq m < n$. *Suppose* $\{T_{m,n}\}$ *has following properties: (i)* $T_{0,n} \leq T_{0,m} + T_{m,n}$. *(ii) The distribution of* $\{T_{m,m+k} : k \geq 1\}$ *does not depend on* m. *(iii)* $\{T_{nk,(n+1)k} : n \geq 0\}$ *is a stationary sequence for each* $k \geq 1$. *(iv)* $E(|T_{0,n}|) < \infty$ *for each* n. *Then, (a)* $\eta = \lim_{n\to\infty} \frac{E(T_{0,n})}{n} = \inf_{n\geq 1} \frac{E(T_{0,n})}{n}$. *(b)* $T = \lim_{n\to\infty} \frac{T_{0,n}}{n}$ *with probability one and* $E(T) = \eta$. *Furthermore, if (v) for* $k \geq 1$, $\{T_{nk,(n+1)k} : n \geq 0\}$ *are ergodic, then (c)* $T = \eta$.

Evidently, if all above conditions are satisfied, Lemma 22 is proved. It is easy to check that condition (i) is satisfied because $T_{0,n}$ is defined as the smallest transmission latency between the SUs located at $\widetilde{X}_m$ and $\widetilde{X}_n$, and it is easy to see $T_{0,n}$ cannot exceed $T_{0,m} + T_{m,n}$, otherwise $T_{0,n}$ is not the smallest transmission latency . Moreover, the conditions (ii) and (iii) are clearly fulfilled because of the stationarity of the homogeneous Poisson point process. Moreover, condition (iv) is met because of Lemma 21. To show that condition (v) of Theorem 23 is satisfied, As in [14], we prove $\{T_{nk,(n+1)k} : n \geq 0\}$ is asymptotically independent, which is a stronger statement than $\{T_{nk,(n+1)k} : n \geq 0\}$ is ergodic.

Lemma 23. *The sequence* $T_{nk,(n+1)k} : n \geq 0$ *is strong mixing.*

Proof. Similar to the proof of Lemma 21, we place two annuli centered at $(nk, 0)$ and $((n + m)k, 0)$, respectively. Since $G(\lambda_s)$ is percolated, by Borel-Cantelli lemma, there always exists an annulus containing an open circuit such that the path with the shortest latency from $\widetilde{X}_{nk}$ to $\widetilde{X}_{(n+1)k}$ is circumscribed within a square S_1 with finite edge length $2^{i+1}d$ and the corresponding path from $\widetilde{X}_{(n+m)k}$ to $\widetilde{X}_{(n+m+1)k}$ is circumscribed within a square S_2 with finite edge length $2^{j+1}d$. As m goes to infinity, the two squares are not overlapping, which means the path from $\widetilde{X}_{nk}$ to $\widetilde{X}_{(n+1)k}$ does not share any links with the path from $\widetilde{X}_{(n+m)k}$ to $\widetilde{X}_{(n+m+1)k}$. Thus, $T_{nk,(n+1)k}$ and

$T_{(n+m)k,(n+m+1)k}$ are asymptotically independent, i.e.,

$$\lim_{m\to\infty} Pr(T_{nk,(n+1)k} < t \cap T_{(n+m)k,(n+m+1)k} < t)$$
$$= Pr(T_{nk,(n+1)k} < t) \cap Pr(T_{(n+m)k,(n+m+1)k} < t)$$

This completes the proof. □

Proof of Theorem 2-2. Without loss of generality, take a straight line passing through X_u and X_v as the x-axis. Consider X_u as the origin. This means $X_u = \widetilde{X}_0$. We denote by n the integer closest to the x-axis coordinate of X_v, which implies that $||X_v - (n,0)|| < \frac{1}{2}$ and $||X_u - X_v|| < n + \frac{1}{2}$

Let d_n be the Euclidean distance between $\widetilde{X}_n$ and $(n,0)$, i.e., $d_n = ||\widetilde{X}_n - (n,0)||$. By triangle inequality, we have $||X_v - \widetilde{X}_n|| < ||\widetilde{X}_n - (n,0)|| + ||X_v - (n,0)|| < d_n + \frac{1}{2} < \infty$. As a result, $T(X_u, X_v) > T_{0,n} - T(X_v, \widetilde{X}_n)$, and thus

$$T(X_u, X_v) = T_{0,n} - \Delta_t, \tag{244}$$

where $\Delta_t < T(X_v, \widetilde{X}_n) < \infty$, which implies

$$\lim_{||X_u - X_v||\to\infty} \frac{T(X_u, X_v)}{||X_u - X_v||} = \lim_{n\to\infty} \frac{T_{0,n}}{n}$$

Thus, from Lemma 5, we obtain

$$Pr(\lim_{||X_u - X_v||\to\infty} \frac{T(X_u, X_v)}{||X_u - X_v||} = \rho) = 1 \tag{245}$$

Now, we derive the upper and lower bounds of ρ. From Proposition 2, ρ is upper bounded by

$$\rho = \inf_{n\geq 1} \frac{E(T_{0,n})}{n} \leq E(T_{0,1}) < \infty \tag{246}$$

To prove the upper bound $\rho > 0$, let $d_0 = ||\widetilde{X}_0 - (n,0)||$ and $d_n = ||\widetilde{X}_n - (n,0)||$ It is clear that $d_0 < \infty$ and $d_n < \infty$. By triangle inequality, we have $||\widetilde{X}_n - \widetilde{X}_0|| > n - d_0 - d_n$. Thus, the number of hops from $\widetilde{X}_n$ to $\widetilde{X}_0$ is at least $\frac{n-d_0-d_n}{r}$. Therefore, ρ is upper bounded by

$$\rho = \inf_{n\geq 1} \frac{E(T_{0,n})}{n} \geq \lim_{n\to\infty} \frac{E(T_i)(n - d_0 - d_n)}{rn} = \frac{E(T_i)}{r} > 0$$

□

6.6 Simulation Results

In this section, we evaluate the impact of node mobility on delay-bounded connectivity through simulations. Consider a secondary network coexisting with a primary network, where primary user density $\lambda_p = 0.1$ and secondary user density $\lambda_s = 1.6$. All primary users have a transmission range $R = 1.2$, while secondary users have a transmission range $r = 1$.

We first study the connectivity of a stationary secondary network with all secondary users being static. In this case, since the SU density $\lambda_s = 1.6$ is larger than the critical one $1.43 < \lambda_c < 1.44$, the secondary network is percolated or connected if the primary network does not exist. As shown in Figure 27, the largest connected component, represented by red dots, contains the majority of secondary users in the network.

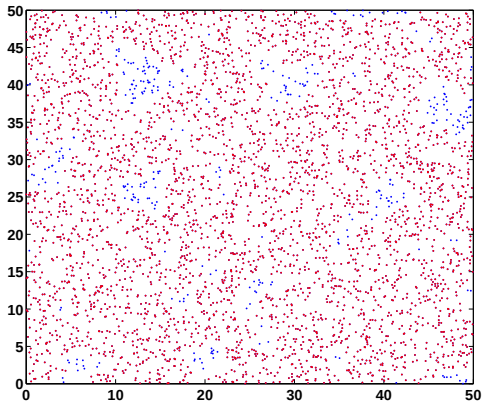

Figure 27: Largest connected component (in red dots) of a standalone secondary network.

However, as indicated by Theorem 19, because of the HT interference from the primary network, there exists a critical density λ_p^* of primary network, above which delay-bounded connectivity is not achievable. According to the network settings, we

have $\lambda_p^* = 0.3025$. In this case, as shown in Figure 28, the largest connected component, represented by red dots, only contains a very small portion of the secondary users, which implies that the whole network is partitioned by the HT interference into small components which are disconnected from each other. Moreover, as implied by Theorem 19, the critical density λ_p^* of primary network does not depend on the density of secondary network, which implies that increasing the density of secondary users cannot improve the connectivity of the secondary network. This can be observed in Figure 29, where the secondary network is still disconnected even if the secondary network in Figure 28 has much higher density $\lambda_s = 5$ than the secondary network in Figure 28.

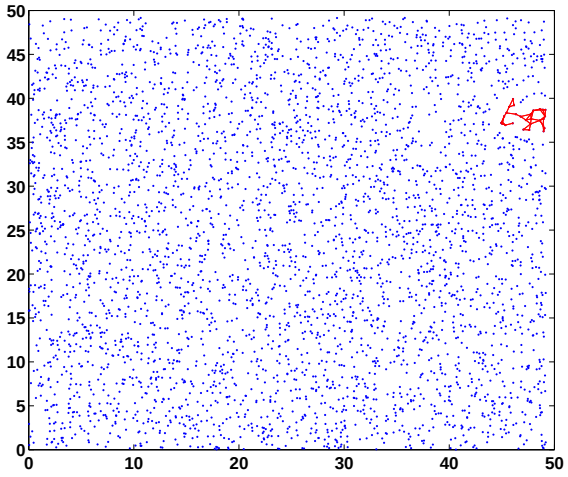

Figure 28: Largest connected component (in red dots) of a secondary network ($\lambda_s = 1.6$) coexisting with a primary network ($\lambda_p = 0.1$).

Next, we investigate the impact of node mobility on the secondary network connectivity. To combat the impact of HT interference of PUs, the spatial diversity of the spectrum can be exploited by allowing mobile secondary users to exchange messages when they opportunistically move into the same white space, the region without PU interference, and are closely enough for data communications. As indicated in Theorem 20, there exists a critical mobility radius a^* above which the secondary network

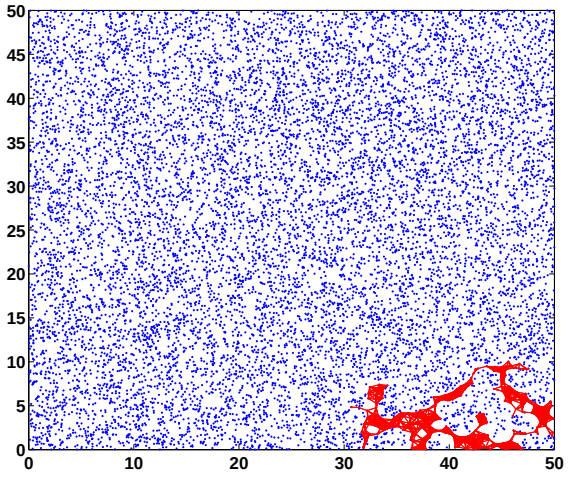

Figure 29: Largest connected component (in red dots) of a secondary network ($\lambda_s = 5$) coexisting with a primary network ($\lambda_p = 0.1$).

can achieve delay-bounded connectivity surely. Based on the network settings at the beginning of this section, we have $a^* = 12.8$. To validate the existence of a^*, we let all secondary users have the maximum mobility radius equal to $a^* = 12.8$ and evaluate the network connectivity in Figure 31 accordingly. It is shown that with the help of node mobility, the secondary network can achieve delay-bounded connectivity so that almost every secondary user resides in a giant connected component. In the contrary, as shown in Figure 30, without node mobility, the same secondary network is disconnected in the sense that the largest connected component only contains a small portion of secondary users because of the HT interference region of PUs, which is denoted by black discs.

We now evaluate the end-to-end average latency in the secondary network with mobile users. More specifically, we assume that all secondary users have their maximum mobility radius larger than the critical one a^* so that delay-bounded connectivity is guaranteed with the rise of a giant connected component in the network. Then, we randomly select a secondary user from this giant component and evaluate the paths with the smallest average latency between this secondary user and every other

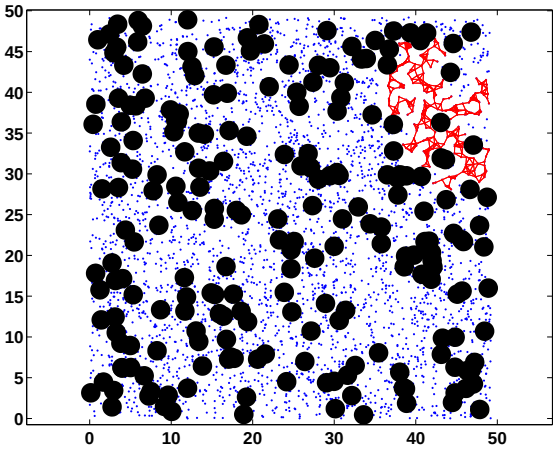

Figure 30: Largest connected component (in red dots) of a secondary network with static users.

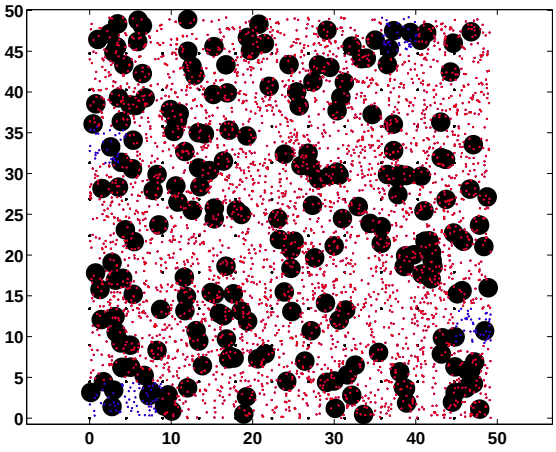

Figure 31: Largest connected component (in red dots) of a secondary network with mobile users.

secondary user it can connect to. As indicated by Theorem 21, the latency along those paths, formally defined as the first passage latency, scales linearly as the initial distance between the sending and receiving parties become large. Such asymptotic linear relationship can be seen in Figure 32, where the ratio between the first passage latency and the initial node distance approaches a constant as the distance increases.

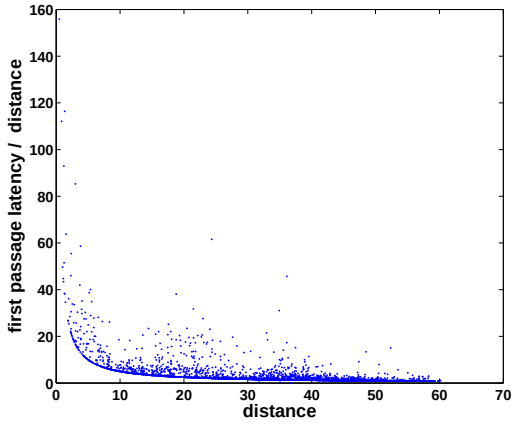

Figure 32: First passage latency of a secondary network with mobile users.

CHAPTER VII

CONCLUSION

7.1 Research Contributions

Recent empirical evidence establishes that the key attributes of wireless networks have exhibited heavy-tailed behavior, which can lead to the extremely bursty nature in Internet and multimedia traffic, the highly variable channel condition, and the irregular mobility pattern of network users. Such heavy-tailed nature largely departs from conventional light-tailed assumptions, and requires the revision of some well-established design principles in communication and network systems. To this end, we first develop novel traffic models that reveal the new origins of heavy-tailed traffic. Then, we analyze the theoretical performance limits in terms of network latency, stability, and connectivity under heavy-tailed network environment. Finally, we propose optimal algorithms spanning different protocol layers to approach these limits.

In Chapter 3, a novel traffic model is proposed, which captures the inherent relationship between heavy-tailed traffic and network dynamics. Then, the statistical attributes of the proposed model are analyzed, which establish the conditions under which user mobility associated with spatial correlation can lead to heavy-tailed traffic. More specifically, it is shown that a high mobility variance and small spatial correlation can give rise to pseudo long range dependent (LRD) traffic, whose autocorrelation function decays slowly and hyperbolically up to a certain cutoff time lag. Secondly, due to the ad-hoc nature of WSNs, certain relay nodes may have several routes passing through them, necessitating local traffic aggregations. At these relay nodes, our model predicts that the aggregated traffic also exhibits the bursty behavior characterized by a scaled power-law decayed autocovariance function. According

to these findings, a novel traffic shaping protocol using movement coordination is proposed to facilitate effective and efficient resource provisioning strategy. Finally, simulation results reveal a close agreement between the traffic pattern predicted by our theoretical model and the simulated transmissions from multiple independent sources, under specific bounds of the observation intervals

In Chapter 4, the asymptotic delay distribution of wireless users is analyzed under different traffic patterns and spectrum conditions, which reveals the critical conditions under which wireless users can experience heavy-tailed delay with significantly degraded QoS performance. More specifically, it is shown that the emerging dynamic spectrum access scheme induces only light-tailed delay if both the busy time of wireless channels and the message size of network users are light-tailed. On the contrary, if either the busy time or the message size is heavy tailed, then the users' transmission delay is heavy tailed. For this latter case, it is proven that if one of either the busy time or the message size is light-tailed and the other is regularly varying with index α, the transmission delay is regularly varying with the same index. As a consequence, the delay has an infinite mean provided $\alpha < 1$ and an infinite variance provided $\alpha < 2$. Furthermore, if both the busy time and the message size are regularly varying with different indices, then the delay tail distribution is as heavy as the one with the smaller index. Moreover, the impact of spectrum mobility and multi-radio diversity on the delay performance of network users is studied. It is shown that both spectrum mobility and dynamic multi-radio diversity can greatly mitigate the heavy tailed delay by maximizing the orders of its finite moments, while by doing the opposite, static multi-radio diversity can aggravate the heavy-tailed delay.

Based on the delay analysis, in Chapter 5, a new network stability criterion, namely moment stability, is introduced to better characterize the stability performance in the presence of heavy-tailed traffic. Then, an asymptotic queueing analysis is performed to reveal the critical conditions under which there exists a feasible

scheduling policy to achieve moment stability. More specifically, it is shown that moment stability is only achievable if the heavy-tailed channel busy time has a tail index larger than three. Utilizing this analysis, a maximum-weight-β scheduling algorithm is proposed, which associates each queue with a different parameter β and makes the scheduling decision based on the queue lengths raised to the β-th power. It is proven that the maximum-weight-β scheduling algorithm is throughput-optimal in the sense that it can maximize the network throughput, while maintaining moment stability. ,

Besides stability, network connectivity also needs to be revisited under heavy-tailed environment. Towards this end, in Chapter 6, a new connectivity criterion, namely delay-bounded connectivity, is introduced, which simultaneously ensures the existence of routing paths and the finiteness of the average delay and jitter along these paths. Then, the sufficient conditions on the existence of delay-bounded connectivity are derived. Specifically, it is proven that if the busy time of primary users is heavy-tail distributed, there always exists a critical density λ_p such that if the density of primary users is larger than λ_p, the secondary networks can not achieve delay-bounded connectivity. To encounter this, the mobility of secondary users is utilized to exploit the spatial diversity of the spectrum availability. In particular, as an important design parameter for all mobility-assisted data forwarding schemes, the critical mobility radius is derived, which is a critical threshold on the maximum radius the network user can reach, above which delay-bounded connectivity is guaranteed. In this case, the end-to-end latency between mobile users is proven to be asymptotically linear in the Euclidean distance between the transmitter and receiver. .

7.2 Future Work

In the future, we intend to broadly revisit wireless networking operations under heavy-tailed environment. One of such operations is multi-hop routing. In particular, we will investigate the effectiveness of the celebrated routing protocols under heavy-tailed

traffic and radio spectrum. For example, it is interesting to investigate whether the widely applied throughout-optimal routing schemes, such as backpressure routing [39], can still maintain their optimality in the presence of heavy tails. Moreover, to avoid network congestion, congestion control solutions need to be developed under the hybrid network traffic, consisting of both inelastic traffic and elastic traffic. Inelastic traffic is generally real-time delay-sensitive traffic (e.g., video and audio traffic) whose data rate can be effectively controlled. In the contrary, elastic traffic (e.g., emails, http traffic, file transfers) is delay-insensitive, whose data rate can be controlled. Because of the strict QoS requirement, inelastic traffic normally has higher priority over elastic traffic. This means the residue bandwidth allowed for elastic traffic to adjust its data rate is fluctuated by the data rate of inelastic traffic. This fact imposes great challenge on the design of effective congestion control solutions because inelastic traffic (e.g., video and audio traffic) normally exhibits heavy-tailed nature with extremely high variability. After investigating the impact of heavy-tailed environment on routing and congestion control, we will develop a light-weight cross-layer framework that incorporates routing and congestion control into our proposed scheduling algorithms, with an objective to maximize network utility in the presence of heavy tails, while incurring limited message exchange and dependency across network layers.

Besides developing effective network control solutions under the generic settings of wireless networks, we will develop effective network management schemes for the emerging Internet of Things based on the theoretical foundations established in this dissertation. Largely departing from the original definition of Internet of Things centered at networked RFIDs, the wireless interconnection of pervasively deployed multimedia devices, ranging from low-cost multimedia sensor nodes to full-fledged smart devices, with existing communication networks and ultimately the Internet defines a truly Cyber-Physical system, which can be referred to as the Internet of Multimedia Things (IoMT). As a major bottleneck of IoMT, the extremely high heterogeneity

in data traffic types, hardware capacities, communication requirements, and operating physical environments necessitates the development of a coherent context-aware networking architecture, which enables devices to intelligently, proactively, and collaboratively adjust their operations according to the heterogeneous context information. To this end, we will continue my current research in the following directions: (I) the impact of mixed heavy-tailed content (e.g., video and internet traffic) and light-tailed content (e.g., temperature and humidity data) on network management; (II) the impact of bursty spectrum on pricing modeling and operator spectrum/revenue sharing schemes. Moreover, I will incorporate two new concepts, i.e., device awareness and social awareness, into the context-aware networking framework. Both concepts can improve the efficient data dissemination by utilizing the heterogeneity in devices, e.g., static low-cost sensor nodes and mobile smart phones, combined with the social behavior patterns of mobile users, including social contacts, social interests, and social relations.

REFERENCES

[1] Akyildiz, I. F. and Kasimoglu, I. H., "Wireless sensor and actor networks: Research challenges," *Ad Hoc Networks Journal (Elsevier)*, vol. 2, no. 4, pp. 351 – 367, 2004.

[2] Akyildiz, I. F., Lee, W. Y., and Chowdhury, K., "Crahns: Cognitive radio ad hoc networks," *Ad Hoc Networks (Elsevier) Journal*, vol. 7, pp. 810–836, July. 2009.

[3] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., and Cayirci, E., "wireless sensor networks: A survey," *Computer Networks (Elsevier) Journal*, vol. 4, pp. 393–422, Mar. 2002.

[4] Bingham, N. H., Goldie, C. M., and Teugels, J. L., *Regular Variation*. Cambridge University Press, 1989.

[5] Butler, Z. and Rus, D., "Event-based motion control for mobile-sensor network," *IEEE Transactions on Pervasive Computing*, vol. 5, no. 3, pp. 34–42, 2003.

[6] Cline, D. B. H. and Hsing, T., "Large deviation probabilities for sums of random variables with heavy or subexponential tails." Technical Report, Texas A & M University.

[7] Cohen, J. W., "Some results on regular variation for distributions in queueing and fluctuation theory," *J. Appl. Probab.*, vol. 10, p. 343C353, 1973.

[8] Cristescu, R., Beferull-Lozano, B., and Vetterli, M., "Networked slepian-wolf: theory, algorithms, and scaling laws," *IEEE Transactions on Information Theory*, vol. 51, pp. 4057–4073, Dec. 2005.

[9] D. Willkomm, S. Machiraju, J. B. and Wolisz, A., "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Comm. Mag.*, vol. 47, no. 3, pp. 88–95, 2009.

[10] Dahlhaus, R., "Efficient parameter estimation for self-similar processes.," *Annals of Statistics*, vol. 17, pp. 1747 – 1766, 1989.

[11] Daley, D. J., "The moment index of minima ii," *Journal of Applied Probability*, vol. 38, no. A, pp. 33– 36, 2001.

[12] Demirkol, I., Alagoz, F., Delic, H., and Ersoy, C., "Wireless sensor networks for intrusion detection: Packet traffic modeling," *IEEE Communication Letters*, vol. 10, no. 1, pp. 22 – 24, 2006.

[13] Demirkol, I., Alagoz, F., Delic, H., and Ersoy, C., "Wireless sensor networks for intrusion detection: packet traffic modeling," *IEEE Communication Letters*, vol. 10, pp. 22–24, Jan. 2007.

[14] Dousse, O., Baccelli, F., and Thiran, P., "Latency of wireless sensor networks with uncoordinated power saving mechanisms," in *In Proc. of ACM MOBIHOC'04*, (Tokyo, Japan), pp. 109–120, 2004.

[15] Dousse, O., Baccelli, F., and Thiran, P., "Impact of interferences on connectivity of ad hoc networks," *IEEE/ACM Transactions on Networking*, vol. 13, pp. 425–436, Apr. 2005.

[16] Duffield, N. G. and O'Connell, N., "Large deviations and overflow probabilities for the general single-server queue,with application," *In Proc.of the Cambridge Philosophy Society*, pp. 363–375, 1995.

[17] Fay, G., Gonzalez-Arevalo, B., Mikosch, T., and Samorodnitsky, G., "Modeling teletraffic arrivals by a poisson cluster process," *Queueing Systems: Theory and Applications*, vol. 4, no. 2, pp. 121–140, 2006.

[18] Geirhofer, S. and Tong, L., "Dynamic spectrum access in the time domain: Modeling and exploiting white space," *IEEE Comm. Mag.*, vol. 45, pp. 66–72, 2007.

[19] Giordano, V., Ballal, P., Lewis, F., Turchiano, B., and Zhang, J., "Supervisory control of mobile sensor networks: math formulation, simulation, and implementation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36, no. 4, pp. 806 – 819, 2006.

[20] Grimmett, G., *Percolation*. New York: Springer, second ed., 1999.

[21] Grossglauser, M. and Bolot, J.-C., "On the relevance of long-range dependence in network traffic," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 629 – 640, Oct. 1999.

[22] Grossglauser, M. and Tse, D., "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 477 – 486, Aug. 2002.

[23] Hwang, C. L. and Li, S. Q., "On input state space reduction and buffer noneffective region," in *Proc. of IEEE INFOCOM'94*, (Toronto, Canada), pp. 1018–1028, 1994.

[24] Jagannathan, K., Markakis, M., Modiano, E., and Tsitsiklis, J. N., "Queue length asymptotics for generalized max-weight scheduling in the presence of heavy-tailed traffic," in *Proc. IEEE INFOCOM*, Apr. 2011.

[25] Kong, Z. and Yeh, E. M., "A distributed energy management algorithm for large-scale wireless sensor networks," in *Proc. of ACM MOBIHOC'07*, (Montreal, Canada), pp. 209 – 218, 2007.

[26] Kong, Z. and Yeh, E. M., "On the latency for information dissemination in mobile wireless networks," in *Proc. of MobiHoc 2008*, May 2008.

[27] Kyasanur, P. and Vaidya, N. H., "Capacity of multi-channel wireless networks: impact of number of channels and interfaces," in *Proc. of MobiCom 2005*, (Cologne, Germany), pp. 43–57, Sept. 2005.

[28] Laourine, A., Chen, S., and Tong, L., "Queuing analysis in multichannel cognitive spectrum access: A large deviation approach," in *Proc. IEEE INFOCOM*, Mar. 2010.

[29] Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V., "On the self similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 204–239, Feb. 1994.

[30] Liggett, T., "An improved subadditive ergodic theorem," *Annals of Prob.*, vol. 13, pp. 1279–1285, 1985.

[31] Lowen, S. B. and Teich, M. C., *Fractal-Based Point Processes*. New York, NY, USA: John Wiley and Sons, Inc., first ed., 2005.

[32] Luo, S., Li, J., Park, K., and Levy, R., "Exploiting heavy-tailed statistics for predictable qos routing in ad hoc wireless networks," in *Proc. of IEEE INFOCOM'08*, (Phoenix, AZ), pp. 2387–2395, Mar. 2008.

[33] Markakis, M. G., Modiano, E. H., and Tsitsiklis, J. N., "Scheduling policies for single-hop networks with heavy-tailed traffic," in *Proc. Allerton Conference on Communication, Control, and Computing*, (Monticello, IL), pp. 112–120, Sept. 2009.

[34] Meester, R. and Roy, R., *Continuum Percolation*. New York, NY, USA: Cambridge University Press, first ed., 1996.

[35] Messier, G. and Finvers, I., "Traffic models for medical wireless sensor networks," *IEEE Communication Letters*, vol. 11, pp. 13–15, Jan. 2007.

[36] Nagaev, S. V., "Large deviations of sums of independent random variables," *The Annals of Probability*, vol. 7, no. 5, pp. 745–789, 1979.

[37] Nair, J., Andreasson, M., Andrew, L. L. H., Low, S. H., and Doyle, J. C., "File fragmentation over an unreliable channel," in *Proc. IEEE INFOCOM*, (San Diego, CA), pp. 1–9, Mar. 2010.

[38] Neely, M. J., "Stability and capacity regions or discrete time queueing networks." Available: `http://arxiv.org/abs/1003.3396v1`.

[39] Neely, M. J., Modiano, E., and Rohrs, C. E., "Dynamic power allocation and routing for time varying wireless networks," *IEEE Signal Processing Magazine*, vol. 23, no. 1, pp. 89 – 103, 2005.

[40] Norros, I., "Large deviations and overflow probabilities for the general single-server queue,with application," *Queueing system*, vol. 16, pp. 387–396, 1994.

[41] Park, K. and Willinger, W., *Self-Similar Network Traffic and Performance Evaluation*. A Wiley-Interscience Pubilication, 2000.

[42] Petriu, E., Whalen, T., Abielmona, R., and Stewart, A., "Robotic sensor agents: a new generation of intelligent agents for complex environment monitoring," *IEEE Magazine on Instrumentation and Measurement*, vol. 7, no. 3, pp. 46–51, 2004.

[43] Ren, W., Zhao, Q., and Swami, A., "Connectivity of heterogeneous wireless networks," *IEEE Transactions on Information Theory*, vol. 57, pp. 399–407, 2011.

[44] Rhee, I., Shin, M., Hong, S., Lee, K., and Chong, S., "On the levy-walk nature of human mobility," in *Proc. of INFOCOM'08*, (Phoenix, AZ), pp. 924–932, 2008.

[45] Rudin, W., *Real and Complex Analysis.* McGraw-Hill, 1987.

[46] Sriram, K. and Whitt, W., "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 833–846, Sept. 1986.

[47] Tan, J. and Shroff, N. B., "Transition from heavy to light tails in retransmission durations," in *Proc. of IEEE INFOCOM'10*, (San Diego, CA), pp. 1334–1342, Mar. 2010.

[48] Tassiulas, L. and Ephremides, A., "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transacations on Automatic Control*, vol. 37, pp. 1936–1949, Dec. 1992.

[49] Tassiulas, L. and Ephremides, A., "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, p. 1936C1948, 1992.

[50] Vuran, M. C., Akan, O. B., and Akyildiz, I. F., "Spatio-temporal correlation: Theory and applications for wireless sensor networks," *Computer Networks (Elsevier) Journal*, vol. 45, no. 3, pp. 245–261, 2006.

[51] Vuran, M. C. and Akyildiz, I. F., "Spatial correlation-based collaborative medium access control in wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 14, pp. 316–329, Apr. 2006.

[52] Wang, P. and Akyildiz, I. F., "Spatial correlation and mobility aware traffic modeling for wireless sensor networks," in *Proc. of IEEE GLOBECOM 2009*, (Honolulu, USA), pp. 1–6, Nov. 2009.

[53] Wang, P. and Akyildiz, I. F., "Dynamic connectivity of cognitive radio ad-hoc networks with time-varying spectral activity," in *Proc. of IEEE GLOBECOM 2010*, (Miami, Florida), pp. 1–5, Mar. 2010.

[54] Wang, P. and Akyildiz, I. F., "Can dynamic spectrum access induce heavy tailed delay?," in *Proc. IEEE DySPAN 2011*, (Aachen,Germany), pp. 197 – 207, May 2011.

[55] Wang, P. and Akyildiz, I. F., "Spatial correlation and mobility aware traffic modeling for wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 19, pp. 1860–1873, Dec. 2011.

[56] Wang, P. and Akyildiz, I. F., "Network stability of dynamic spectrum access networks in the presence of heavy tailed traffic," in *Proc. IEEE SECON 2012*, (Seoul, Korea), pp. 165 – 173, 2012.

[57] Wang, P. and Akyildiz, I. F., "On the origins of heavy tailed delay in dynamic spectrum access networks," *IEEE Transactions on Mobile Computing*, vol. 11, no. 2, pp. 204–217, 2012.

[58] Wang, P. and Akyildiz, I. F., "Asymptotic queueing analysis for dynamic spectrum access networks in the presence of heavy tails," *IEEE Journal on Selected Areas in Communications*, vol. 3, no. 3, pp. 514–522, 2013.

[59] Wang, P., Akyildiz, I. F., and Al-Dhelaan, A. M., "Mobility improves delay-bounded connectivity in dynamic spectrum access networks," *Submitted for journal publication.*

[60] Wang, P., Akyildiz, I. F., and Al-Dhelaan, A. M., "Percolation theory based connectivity and latency analysis of cognitive radio ad hoc networks," *ACM Journal of Wireless Networks*, vol. 17, no. 3, pp. 659 – 669, 2011.

[61] Wang, S., Zhang, J., and Tong, L., "Delay analysis for cognitive radio networks with random access: A fluid queue view," in *Proc. IEEE INFOCOM*, Mar. 2010.

[62] Wellens, M. and Mahonen, P., "Lessons learned from an extensive spectrum occupancy measurement campaign and a stochastic duty cycle model," *Mobile Networks and Applications*, vol. 15, no. 3, pp. 461–474, 2010.

[63] Wellens, M., Riihijarvi, J., and Mahonen, P., "Empirical time and frequency domain models of spectrum use," *Physical Communication*, vol. 2, pp. 10–32, Mar. 2009.

VITA

Pu Wang received the B.S. degree in Electrical Engineering from the Beijing Institute of Technology, Beijing, China, in 2003 and the M.Eng. degree in Com- puter Engineering from the Memorial University of Newfoundland, St. Johns, NL, Canada, in 2008. He is now a PhD candidate in the school of Electrical and Computer Engineering of the Georgia Institute of Technology, Atlanta, GA, in 2013. Starting August 2013, he will be an Assistant Professor with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS. He was named BWN Lab Researcher of the Year 2012, Georgia Institute of Technology. He received the TPC top ranked paper award of IEEE DySPAN 2011. He was also named Fellow of the School of Graduate Studies, 2008, Memorial Univer- sity of Newfoundland. He is a member of IEEE. His research interests include wireless sensor networks, cognitive radio networks, nanonetworks, multimedia communications, wireless communications in challenged environment, Internet of things, and cyber-physical systems