Analyzing the Signal Strength of 2,946 Clients Operating in 446 WiFi Networks

by

Midul Kurian Jacob

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2020

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis we analyze data that was collected over a 24 hour period from 446 access points that provide connections for 2,946 clients. The data was obtained from deployments of modern commercial Google Wifi access points. A focus of this thesis is an analysis of the Received Signal Strength Indication (RSSI) from messages sent from clients to the access point. The RSSI depends on both the environment in which the signals operate and the distance between the client and the access point.

Our dataset includes 417,122 data points of which 45.1% of the data points are from signals using the 2.4 GHz spectrum and the remaining 54.9% are from the 5 GHz spectrum. The data has been collected by each access point (AP) every 5 minutes over a period of 24 hours. We find from our analysis that across all access points, the average number of clients (across all spectrums) that are simultaneously connected in any 5 minute window is quite small. That is, 65.7% of the APs have on average 3 or fewer clients that are simultaneously connected in any 5 minute window. However, we also find that 6.5% of the APs service on average 9 or more clients.

In this thesis we develop and utilize a methodology to categorize clients and networks using RSSI values (signal strengths) of the messages received by access points from the clients, to study the possible PHY rates which can be used by clients to send messages to the APs. The methodology also helps us to capture and examine the variability in signal strengths.

Several previous studies have characterized WiFi networks using the measured throughput of the clients. However, the throughput experienced and rates used by clients in those studies depend on the capabilities of the clients. We believe that a significant advantage of our methodology is that it is independent of the capabilities of the clients used in the study. In addition, our methodology is also able characterize the environments in which the WiFi devices operate. This is because our methodology primarily uses the signal strength of the messages to characterize devices in a WiFi network and the signal strength changes over time due to the movements of the sender, receiver, or people in the area.

We use our methodology to analyze both clients and networks. We find from our analysis of clients that, over the 24 hour period, 90% of the signals from 84.2% of the clients are received by the APs with either Good or Moderate signal strengths. Thus, for the majority of the clients signal strengths are mostly quite reliable. We also find that clients using the 2.4 GHz spectrum have signals about as good as or better than clients using the 5 GHz spectrum.

However, perhaps one of the most interesting findings is that, when analyzing networks we find that 27% or more of all networks have one or more clients whose signals are received by the AP with unreliable signal strengths. These clients could potentially impede the throughput of all the other clients in the same network and also networks in the vicinity, due to the WiFi performance anomaly problem. We also find that networks with more clients have more clients with unreliable signals and that the fraction of networks with one or more clients with unreliable signals is quite close to what is expected based on probabilities. From the results of our analysis

of clients and the analysis of networks, we note that a small number of clients may impact the performance of a considerably large number of networks.

Acknowledgements

I would like to express my most sincere gratitude to my supervisor, Tim Brecht, for his continuous support and guidance in all stages of my research; both intellectually and financially. His consistently detailed feedback and research ideas helped me immensely in my graduate studies and in the successful completion of this thesis. Therefore, I sincerely thank him and shall always cherish his sense of humor, enthusiasm and encouragement in the most challenging of circumstances.

Furthermore, I thank my thesis readers, Ali Abedi and Martin Karsten, for their insightful feedback that helped improve this thesis.

I am also grateful for my friends who stood by me, encouraged me, and above all, had faith in me when mine faltered. Sincere thanks to Rafael Toledo, David Thomas Radke, Andrea D'Alessandro, Erin Kim, Michael Milne, Jason Daniel, Khristen Lowe, Sultana Sharmin and Shervin Khastoo. Thank you for being there.

Additionally, I also thank God for giving me the knowledge and intellect to persevere and complete my graduate thesis successfully.

I express my appreciation to my girlfriend Susanna George for all her love and support. Thank you for sticking with me and encouraging me during the long months of writing and re-writing this thesis.

Finally, I thank my parents, Jacob Kurian and Valsa Jacob, and my brother, Jidul Jacob, for all their sacrifices, constant support and prayers throughout this arduous journey. I dedicate this thesis to them.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The 802.11n and 802.11ac standards enable WiFi communication between two wireless devices, a sender and a receiver. For effective communication, messages from the sender must arrive at the receiver with sufficient signal strength so that they can be decoded. In this thesis, we present an analysis of anonymized data provided by Google that was obtained from randomly sampled commercial Google Wifi access points (APs). We analyze and study data obtained from the access points and categorize clients and networks using the signal strength measured by the AP as reported by the Received Signal Strength Indication (RSSI).

1.1 Motivation

Advances in communication technology and the proliferation of smart devices in personal, public and business spaces have seen access to wireless networks increase at a dramatic rate. The WiFi market is estimated to grow from 5.96 billion USD in 2017 to 15.60 billion USD by 2022. The worldwide market for WiFi home routers will reach 3.57 billion USD in 2024, up from 2.67 billion USD in 2019 [15]. In 2018 there were 169 million WiFi hotspots globally, but Cisco expects this to increase to 628 million by 2023 [7]. Cisco also predicts that by 2022 WiFi and mobile networks will account for 71% of IP traffic and that the average WiFi speeds will grow from 30.3 Mbps in 2018 to 92 Mbps by 2023 [33].

What makes WiFi a popular choice is that even with changes in data costs, cellular networks remain limited in spectrum, cost 10 times more to deploy [17] and are often billed on a per-byte basis. WiFi remains cheaper to deploy, provides faster operating speeds and in almost all cases, is free to the end user. While cellular connections work well outside, they often do not work well as users enter the core of many buildings. WiFi networks, on the other hand, can be designed to work consistently throughout internal spaces.

Previous work by Zhou et al. [55] has shown that the throughput offered by internet service providers (ISPs) typically exceeds the throughput available through a WiFi access point. Factors like the signal strength of the received signal, the spectrum and the channel width used in transmission and the attenuation experienced by the signal in the medium of transmission, can potentially limit the range of physical rates that an be used during transmission. Hence, the throughput experienced by a client on a WiFi network is usually dictated by the factors in the environment of the network deployment.

1.2 Goals of the Thesis

An important aspect of WiFi networks that can impact performance is the maximum speed of the physical rates that can be used for transmission. The maximum rates that a client can use for transmission depends on the range of PHY transmission rates supported by the client. For example, if a client only supports 1 spatial stream, that limits the maximum throughput of that client. Likewise, if the client does not support 80 MHz channel widths in the 5 GHz spectrum, then its throughput may potentially be lower than those that do support 80 MHz channel widths. However, when using a fast rate (high PHY rate), if the signal strength of the messages received by the receiver is not strong enough, then the receiver will not be able to successfully decode it and the transmitting client's rate adaptation algorithm (RAA) will infer from packet losses that it must use lower PHY rates. Thus, an important factor in being able to use a particular rate is the ability of the receiver to successfully decode the messages received at a specific signal strength (more details are provided in Section 2.1). Thus, one of the goals of this thesis is to analyze both the clients and the access points by examining the signal strength of the messages received by the AP from the clients.

The focus of this thesis is to characterize the signal strength of 2,946 clients operating in 446 networks. To effectively characterize the signal strength we want to be able to categorize signals from clients as either, Good, Moderate, Variable, or Weak. For instance, we would like to determine the percentage of clients for which the APs receive Good or Weak signals. Similarly, based on the signal strengths of the messages from all the clients at the AP, we also want to distinguish between APs that receive Good, Moderate, Variable and Weak signals. For example, we would like to determine the percentage of networks with clients whose messages are received with Good signals and also identify the percentage of unreliable clients in a network, that is clients whose messages are received by the AP with neither Good nor Moderate signal strengths. An advantage of analyzing signal strengths rather than observed throughput is that our analysis is independent of the particular PHY transmission rates and number of spatial streams supported by the client. As a result, we are able to characterize the environments in which the devices operate.

1.3 Contributions

In this thesis we conduct a detailed analysis of data sampled every 5 minutes over a 24 hour period from 446 commercial Google Wifi access points that provide connections for 2,946 clients.

These devices support the 802.11ac standard and are backwards compatible with the 802.11n standard. All data is anonymized and cannot be traced back to any particular access point, client device or user. Furthermore, no payload data or IP addresses are collected or examined. In this thesis a network refers to an access point and the clients connected to it. The dataset does not include any mesh networks so every network consists of only one access point.

The contributions of this thesis are as follows:

- We study the signal strengths of the messages received from 2,946 clients operating in 446 networks using a total of 417,122 data points. We use RSSI as a measure of the quality of the signal.

- We develop a methodology to categorize clients and APs based on the strength of the signal received by the AP from the client. This provides insights into the range of rates that can be used by the clients for transmission. Our approach to categorization is aimed at better understanding the clients and the networks in terms of the signal strength between them, based on the clients' potential to use high rates. The categories also reflect the characteristics of the environment since the environment affects signal propagation.

- We study the variability in the signal strengths of the signals from the clients using different dispersion thresholds and by categorizing both the clients and the APs into one of four distinct categories: Good, Moderate, Weak and Variable.

- We find from our analysis of clients that, over the 24 hour period, 90% of the signals from 84.2% of the clients are received by the APs with either Good or Moderate signal strength. Thus, for the vast majority of the clients, signal strengths are mostly quite reliable. But we also find that there are a small number of clients that experience Weak signals for most of the time. Finally, we also find that clients using the 2.4 GHz spectrum have signals about as good as or better than clients using the 5 GHz spectrum.

- We find from our analysis of networks that more than 27% of all networks have one or more clients with unreliable signals. This implies that there are a considerable number of networks where the performance may be significantly hindered, due to the WiFi anomaly problem, because one or more clients are operating with unreliable signals and as a result are forced to use slower rates. We also analyze, whether the fraction of networks with unreliable clients is higher than the probabilities would suggest for networks with more clients. We find that networks with more clients have more clients with unreliable signals and that the fraction of networks with one or more clients with unreliable signals is quite close to what is expected based on probabilities. From the results of our analysis of clients and networks, we note that a small number of clients may impact the performance of a considerably large number of networks. The importance of our approach lies in that we examine the health of the entire network.

1.4 Thesis Organization

In Chapter 2 we describe the received signal strength indication (i.e., the signal strength), the physical rate and the different factors that affect the throughput at which a signal can be received. We also review the previous studies that are closely related to the work we do in this thesis. In Chapter 3, we analyze our dataset on the basis of the number of datapoints and the distributions of the minimum, maximum and average number of clients that are connected to the access points over the 24 hour period. We also analyze the number of datapoints and the distribution of the RSSIs of messages received by the APs from all the clients, based on their spectrums and channel widths. In Chapter 4, we devise a methodology to categorize the clients and the networks using the signal strength and to analyze the variability in the signal strengths using dispersion thresholds. We then provide an example categorization of a client and a network using the methodology for different dispersion thresholds. In Chapter 5, we use our methodology to analyze the clients. In Chapter 6, we use our methodology to analyze the networks. Due to the WiFi anomaly problem, we also study networks with one or more unreliable clients. Lastly, we briefly summarize the work presented in this thesis, propose future areas of research and provide concluding remarks in Chapter 7,

Chapter 2

Background and Related Work

2.1 Background

In this section, we provide an overview of the factors that impact the signal strength and the throughput of devices in WiFi networks. We focus on the aspects of these concepts that help us to categorize and study both the clients and the networks based on the signal strength of the messages received at the access point and its ability to decode messages transmitted using different rates.

2.1.1 Physical Transmission

The 801.11n and the 802.11ac standards define protocols for wireless local area networking on both the 2.4 GHz and the 5 GHz spectrums. Communication between any two devices on the WiFi network requires one of the devices to send a message encoded in an analog signal and another device to receive the signal with sufficient signal strength (RSSI) to decode the signal and recover the message. However, the signal strength of the messages arriving at the receiver is affected by the attenuation of the signal in the medium of transmission. Further, the receiver's ability to decode a message depends on the spectrum being used (2.4 GHz or 5 GHz), modulation scheme [26], the coding rate [27, 49], the channel width used for transmission [20] and the sensitivity of the WiFi chipset used in the receiver [31, 48].

2.1.2 Received Signal Strength Indication

The Received Signal Strength Indication (RSSI) is a measure of the strength of the signal when received by a device. Since the 802.11 standard does not specify how the RSSI should be reported, chip manufacturers define their own interpretation of RSSI. However, the reported value of RSSI is always negative. For instance, Cisco defines its RSSIs in the range between 0 and

-100 dBm [32], whereas Atheros defines its RSSIs between 0 and -127 dBm [32]. The closer a value is to zero the higher the RSSI, thereby, indicating a stronger signal. Similarly, a lower RSSI indicates a weaker signal.

RSSIs are expressed in decibel-milliwatts (dBm) using a logarithmic scale [8]. The decibel-milliwatts is the decibel level relative to one milliwatt (mW). The logarithmic nature of dBm can be understood by realizing that a 3 dBm decrease in signal strength halves the signal strength and a 10 dBm decrease indicates a 10 fold decrease in signal strength.

2.1.3 Signal Attenuation

A major problem in WiFi networks is signal attenuation, which is the gradual loss of signal strength as it passes through a medium. This loss can happen for two reasons. First, as signals pass through various obstacles, they are reflected, refracted and absorbed by the obstacles to varying degrees depending on the attenuation factor of the material the object is made of [30, 43, 42].

Second, different frequencies attenuate at different rates. As a result, signal strength is inversely proportional to both the square of the signal frequency and the distance between the sender and the receiver [50]. Thus, if either the frequency of the signal or the distance between the sender and receiver were to be doubled, then the signal strength would decrease by a factor of four. The loss in strength of a signal as it travels through a medium can be represented mathematically as shown in Equation 2.1 where P_l is the average propagation path loss, c is the speed of light, d is the distance between the transmitter and the receiver, f is the frequency at which the signal is transmitted and n is the attenuation exponent, which is 2 for free space [38, 40].

$$P_l \propto (\frac{4\pi f}{c})^2 (d)^n \tag{2.1}$$

2.1.4 Modulation Scheme

Modulation is the process of converting a digital bit stream to an analog signal that can be physically transmitted. Modulation schemes specify the complexity of the modulation technique [19]. The more complex the modulation scheme, the higher the data rate. However, stronger signal strengths are required to decode the information transmitted using more complex modulation schemes.

2.1.5 Coding Rate

To help detect and correct errors in signal transmission, redundant bits are added to the data. The coding rate specifies how much of the transmitted data contains usable information [27, 49]. It is expressed as a ratio of the number of data bits transmitted to the total number of coded bits.

For example, a coding rate of 1/2 implies that one data bit is transmitted for every two code bits. More aggressive codes have less redundant information. For a coding rate of 3/4 only 25% of the bits are redundant. As the coding rate decreases more coding bits are available to correct errors and the code becomes more robust. However, the price of robustness is decreased throughput [16].

The Modulation and Coding Scheme (MCS) index (MCSI) identifies the combination of the modulation scheme and the coding rate used in a particular transmission. For example, the highest MCSI in 802.11ac networks is 9 and it indicates that signals using an MCSI of 9 uses a modulation scheme of 256-QAM (Quadrature Imposture Modulation) with a coding rate of 5/6 (indicating that 83.3% of the data is usable). We do not elaborate further on various modulation and coding techniques, as it is beyond the scope of this thesis.

2.1.6 Channel Width

To increase the data rate of transmissions the 802.11n standard introduced the ability to increase the channel width by combining more than one channel. For example, a 40 MHz channel width combines two 20 MHz channels. Similarly, 80 MHz and 160 MHz channels have been added to the 802.11ac standard. While it is possible to achieve higher data rates by using wider channels, studies have shown that doubling the channel width increases the noise floor by about 3 dB [9, 10]. This not only reduces the transmission range, but also makes signals more sensitive to interference [5]. Thus, as channel width increases, signals must be received with higher signal strengths in order to be correctly decoded.

2.1.7 Multiple-Input/Multiple-Output (MIMO)

MIMO, is an acronym for Multiple-Input/Multiple-Output and is the most important feature introduced in the 802.11n standard, since the physical layer rates can be increased by up to four times using MIMO [3, 24]. In contrast to previous 802.11 standards (i.e., 802.11a, 802.11b and 802.11g), 802.11n and 802.11ac standards utilize multiple antennas for transmission and reception of the data streams. MIMO utilizes a technique called spatial multiplexing to transmit multiple independent streams of data concurrently. This leads to increased PHY rates.

Spatial multiplexing takes advantage of the random nature of signal propagation in environments [4, 51]. This has enabled MIMO technology to be used to transmit multiple streams of data simultaneously, using multiple antennas, over the same wireless channel and successfully decode and merge the streams of data at the receiver. The number of data streams used in transmission and reception is often called spatial streams. More spatial streams provide for higher PHY rates. However, irrespective of the number of data streams being used, each stream of data must be received at the receiver's antenna with sufficient signal strength for the data in the stream to be decoded. Our methodology does not depend on the number of spatial streams supported by the client or the AP, rather it only takes into account if the message received at the AP were of sufficient strength to be decoded.

2.1.8 Guard Interval

Another factor affecting the data rates of WiFi networks is the guard interval. Guard intervals are used to ensure that distinct transmissions do not interfere with one another [37, 52]. These transmissions may belong to different users or to the same user (due to multiple spatial streams). The purpose of the guard interval is to provide immunity to signals from factors like the propagation delays, echoes and reflections. A low guard interval results in a slightly higher data rate.

2.1.9 Noise Floor

In signal theory, noise is defined as any signal in the environment other than the one being monitored. Noise is always present and received on a radio even when no wanted signals are present. Consequently, the noise floor is the measure of the noise from a number of sources including thermal noise, atmospheric noise and noise from components used to make the device. The noise floor is an important aspect of any receiver as it gives an indication of the minimum signal strength required to successfully isolate the signal of interest from the unwanted signals (i.e., the noise) in the environment.

2.1.10 Receiver Sensitivity

The receiver sensitivity is the minimum signal strength required to decode signals and convert them to valid data [31, 48]. The receiver sensitivity depends on the MCS index, the spectrum used for transmission and the channel width. It can usually be obtained from the datasheets provided by the manufacturer for each chipset. In most cases, as the spectrum, the channel width and the MCSI increase a stronger signal strength is required to decode the signal. For the purposes of this study, since the commercial Google Wifi access points used in our study use Qualcomm's IPQ-4019 chipset we use the receiver sensitivity values from Qualcomm's IPQ-4019 [23]. Note however that our methodology can be used for any chipset by simply using the receiver sensitivity from the manufacturer's datasheet and adjusting the RSSI levels used (e.g., Table 4.1).

2.1.11 Factors Affecting PHY Rates

The physical transmission rate (PHY rate) is determined by the combination of modulation technique, coding rate, number of spatial streams, guard interval and channel width. The MCSI (MCS Index) summarizes the combination of the modulation technique and the coding rate used in a particular transmission.

Table 2.1 shows a subset of possible data rates for 802.11ac networks. We observe that as the number of spatial streams, MCSI and channel width increases the PHY rate also increases. In contrast, as the guard intervals increase the PHY rates decreases slightly. This table shows how these different factors affect the PHY rates of devices in a WiFi network.

MCSI	Spatial Streams	Modulation Technique	Coding Rate	Data rate (in Mbit/s)					
				20 MHz Channels		40 MHz Channels		80 MHz Channels	
				800 ns GI	400 ns GI	800 ns GI	400 ns GI	800 ns GI	400 ns GI
0	1	BPSK	1/2	6.5	7.2	13.5	15.0	29.3	32.5
1	1	QPSK	1/2	13.0	14.4	27.0	30.0	58.5	65.0
2	1	QPSK	3/4	19.5	21.7	40.5	45.0	87.8	97.5
3	1	16-QAM	1/2	26.0	28.9	54.0	60.0	117.0	130.0
4	1	16-QAM	3/4	39.0	43.3	81.0	90.0	175.5	195.0
5	1	64-QAM	2/3	52.0	57.8	108.0	120.0	234.0	260.0
6	1	64-QAM	3/4	58.5	65.0	121.5	135.0	263.3	292.5
7	1	64-QAM	5/6	65.0	72.2	135.0	150.0	292.5	325.0
8	1	256-QAM	3/4	78.0	86.7	162.0	180.0	351.0	390.0
9	1	256-QAM	5/6	N/A	N/A	180.0	200.0	390.0	433.3
0	2	BPSK	1/2	13.0	14.4	27.0	30.0	58.5	65.0
1	2	QPSK	1/2	26.0	28.9	54.0	60.0	117.0	130.0
2	2	QPSK	3/4	39.0	43.3	81.0	90.0	175.5	195.0
3	2	16-QAM	1/2	52.0	57.8	108.0	120.0	234.0	260.0
4	2	16-QAM	3/4	78.0	86.7	162.0	180.0	351.0	390.0
5	2	64-QAM	2/3	104.0	115.6	216.0	240.0	468.0	520.0
6	2	64-QAM	3/4	117.0	130.3	243.0	270.0	526.5	585.0
7	2	64-QAM	5/6	130.0	144.4	270.0	300.0	585.0	650.0
8	2	256-QAM	3/4	156.0	173.3	324.0	360.0	702.0	780.0
9	2	256-QAM	5/6	N/A	N/A	360.0	400.0	780.0	866.7

Table 2.1: Examples of some PHY rates for 802.11ac networks

2.1.12 Medium Access Control

In 802.11 medium access control (MAC) sub-layer is responsible for coordinating access to the shared physical air interface so that the Access Point (AP) and client can communicate effectively. MAC implements the control mechanisms that allow multiple devices to reliably communicate by sharing the medium. 802.11 standard uses the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol to share the radio channel in a fair way among all its users. To do this, the transmitting device first senses the medium (air interface in our case). Only if the channel is free for a specified amount of time, can they transmit. If channel is sensed busy, the device waits till it is free for a specified duration and further waits for a random interval of time before transmitting. This randomness reduces the chances that two devices, each competing to transmit, end up transmitting at the same time. CSMA/CA channel access method guarantees that the long term channel access probability is equal for all devices. This also results in the WiFi anomaly problem discussed in the next subsection.

2.1.13 The WiFi Performance Anomaly Problem

WiFi networks also suffer from the performance anomaly problem [25] where clients transmitting using a low data rate affect other clients on the same network. Because all wireless signals are transmitted through the air, which is a shared medium, only one device can transmit at any given time. If one client is restricted to sending messages to the access point using low rates (e.g., because the signal received by the AP is not strong enough to decode higher rates), it occupies the medium for a longer period of time than if it was able to use high rates. For example, a device transmitting at 1 Mbps captures the channel for eleven times longer than devices transmitting at 11 Mbps. As a result, other devices in the vicinity must wait longer for access to the medium (channel). This waiting reduces the effective throughput of all the devices in the vicinity. Consequently, it is critical to understand which clients and how many clients are unable to use higher rates. This is a key focus of our analysis in this thesis.

2.2 Related Work

There has been significant interest in the characterization of WiFi networks. In this section we describe studies that characterize wireless networks using RSSI values and throughput. We also explore studies that use different techniques and criteria to categorize clients and networks. We first summarize the previous studies that are most closely related to the work in this thesis. Then we describe the aspects of our research that differs from that work. We focus mainly on these studies because we primarily study the access points (APs) and clients based on the RSSI values of the signals received by the AP and categorize them based on their potential to use different PHY rates.

2.2.1 Characterizations using RSSIs

In this section we review studies that characterize wireless networks using RSSI values under different settings. We do this to better understand the impact that different factors (e.g., spectrum and distance) have on the RSSI value and to also understand how different RSSI values affect user's wireless experience (e.g., throughput). Two of the studies described below [13, 38] are based on 802.15.4 networks. Devices based on the 802.11 and 802.15.4 standard have similar operating frequencies in the 2.4 GHz spectrum. Therefore, it is reasonable to expect similar radio signal propagation models between the different specifications. Our interest in the studies that analyze 802.15.4 networks is to understand the signal propagation mode and more precisely how the signal strength is affected by various factors in the environment.

Pu et al. [38], in their 2007 paper, study RSSI variations in an indoor setting using 802.15.4 compliant hardware and the 2.4 GHz spectrum, for modelling a system to convert RSSI values to distance to design fine-grained location estimation systems. We summarize the paper here because they find that the RSSI values change with both distance and time.

To study RSSI values, they first measure the raw RSSI values and convert them to the actual received power in dBm by adding an RSSI offset value. The RSSI offset value is empirically determined using the front-end gain by the receiver. Front-end gain measures how how well the receiving antennas' front-end circuitry converts the signal arriving from a specified direction into electrical power. A mobile receiver is placed in an arbitrary location (up to 2.5 m from the transmitter) in an indoor environment (which contains furniture) where the walls are made of concrete and glass. The study analyzes the relationship between the received signal strength and the distance between the transmitter and the receiver.

They find that the RSSIs depend on both space (distance) and time. They find that the RSSI values decrease with increasing distance from the transmitter and that they also vary over time with no changes in distance. They therefore conclude that RSSI alone cannot be used for accurate indoor location estimation and suggest that characterization of any environment using RSSI values should consider both temporal and spatial factors. In other words, they suggest that studies that use RSSI values to characterize an environment should analyze how RSSIs vary over both distance and time. In this thesis, we analyze the variability of the RSSIs using the methodology devised by us in Chapter 4.

Similar to the study by Pu et al. [38], Fang et al. in their 2010 paper [13], study fundamental factors contributing to RSSI variability, from the standpoint of RSSI localization algorithms, using a 802.15.4 compliant transmitter and receiver operating in the 2.4 GHz spectrum. We summarize the paper here because they also analyze how RSSIs are affected in an indoor environment.

To study the change in RSSI with distance, they use one transmitter and measure the RSSI value at the receiver over varying distances in two different scenarios: a) in a large open field (using distances up to 110 m), b) inside a lab hall (using distances up to 70 m).

From the results of the experiments in the open field, they find a perfect logarithmic fit between RSSI and distance and the degree of fit is up to 98.5% when compared with the theoretical model of signal attenuation in open space (e.g., using Equation 2.1). However, from the results of the experiments inside a lab hall, they find large variations in the RSSI values for varying distances and that there is no observable pattern. Unlike Pu et al. [38] they do not inspect if RSSIs change with time without changing the distance between the transmitter and the receiver. Their results show that RSSI can work well for localization techniques only for limited scenarios, such as an ideal open outdoor environment. In other scenarios, such as a large room, RSSI values are affected by reflecting and attenuating objects in the environment. Therefore, from the results of [13, 38] we understand that analyzing the variability in RSSIs can help characterize the environments in which the signals operate.

Tauber et al. in their 2007 paper [46], study the impact of low RSSIs on throughput in 802.11n and 802.11g networks. They do this since low signal strengths can impact the performance of applications by affecting the throughput. This is because for a given channel width and spectrum the RSSI values limit the range of MCSIs that a transmitted signal can be encoded with for it to be decoded successfully by the receiver. These MCSIs determine the maximum PHY rates that can be used for transmission.

They evaluate the impact of low RSSIs using a WLAN testbed capable of supporting 802.11g and 802.11n networks. The experiment was set up in a university teaching lab with a distance of 24 m between the transmitter and the receiver for (in their words) good RSSIs. To emulate poor RSSIs they use 10 dB attenuators to weaken the signal. They emulate traffic flows using `iperf` [47] and collect throughput measurements. To allow for comparisons between 802.11g and 802.11n networks they constrain the upper limit of offered throughput to the upper limit of throughput expected in 802.11g networks (in their case this is 30 Mbps).

They find that in an office environment with a clear line of sight (i.e., with good RSSI) between the transmitter and the receiver, the 802.11n network reaches higher throughput (up to 30 Mbps) compared to 802.11g networks (up to 20 Mbps). However, they also find that under poor RSSI conditions throughput is reduced to about 2.5 Mbps in both the 802.11n and 802.11g networks and that hardly any difference exists between them. Therefore, poor RSSIs can have significant impact on the throughput experienced by the clients operating in 802.11n and 802.11g networks.

Azini et al. in their 2015 paper [1], study antenna design for WiFi applications. They also study the relationship between different RSSI values and throughput in the 2.4 GHz spectrum by categorizing the RSSI values. While their primary goal is to analyze the antenna performance, we summarize the paper here because, like us, they divide the RSSI values into different categories.

The antenna they use to transmit signals to the receiver connects to an R52 radio module, which supports 802.11a, 802.11b and 802.11g standards. The RSSI values were measured using an AirMagnet wireless PC card that supports 802.11n networks and is backward compatible with 802.11a, 802.11b and 802.11g standards. The throughput is measured using WinBox Loader on the PC. The RSSI measurements were taken for three different scenarios: transmission within a hall, transmission within a room and transmission outside the room. The RSSI measurements are classified into five categories: Excellent ($\geq$ -60 dBm), Good (-61 dBm to -75 dBm), Fair (-76 dBm to -80 dBm), Bad (-81 dBm to -89 dBm) and Very Bad ($\leq$ -90 dBm). Their categories are based on the signal strength indication of the AirMagnet wireless PC Card and past studies [12, 29]. We describe these studies later in more detail in Section 2.2.3.

Their investigations reveal that the average throughput (over 10 secs) recorded by the client is proportional to different RSSI categories. In other words, they find that the client records higher throughput when the RSSI value belongs to a better category (e.g., the Good category is better than Fair).

Habaebi et al. [22] in their 2016 paper, study whether they can leverage the variability in RSSI values in WiFi networks to develop control systems. An example would be a system that turns on the lights in an office lounge when there are variations in the RSSI values due to the movement of people in the room (e.g., when someone enters the lounge).

The authors leverage the understanding that due to the effects of multipath, signals arriving at a receiver in an environment where people are moving will have variable signal strengths. This is because, as people move around in an environment, signals reflect and refract off of them in different directions at different times. Consequently, the signals reach the receiver by two or more paths. While in some cases this can help improve the signal strength, at other times it can deteriorate the signal strength. This results in variability in signal strengths.

Knowing this, the authors analyze the variability in RSSIs between an empty and a non-empty room in a university. An empty room is one without any people present. A non-empty room is one where there is some presence and movement of people in the environment. For this the authors use two different environments: a dormitory room and a university staff room. In these environments the presence of a person and their movement acts as an obstacle for the signals. The RSSI values are recorded using a receiver that supports 802.11b, 802.11g and 802.11n standards. The distance between the receiver and the transmitter is 3 m.

They find that the variance of the RSSIs in a non-empty room is significantly higher than that in an empty room. Further, RSSI values from the non empty room have a weaker maximum and minimum than from an empty room. Therefore, by analyzing the variability in the RSSIs they can characterize the environment of signal propagation. That is, if the variability in the RSSI values is high, then the device may be operating in an environment where there are people moving within the room. Moreover, we know from the results of Tauber et al. [46] that signals with weaker RSSIs have lower throughput. Therefore, since the movement of people also weakens the maximum and minimum RSSI values, the movement of people may also reduce throughput.

Similar to the study in 2016 by Habaebi et al. [22], Rosli et al. in their 2018 paper [39] explore whether RSSIs can be used for different IoT applications and their implementations. An example would be a security system that triggers an alert when someone enters a room. They study RSSI values from two aspects. First, they study how RSSI values change in the presence of an obstacle (i.e., a 75 mm thick brick wall). Second, they analyze how human movement between the transmitter and receiver that are placed 1 m apart affects RSSI values. Their transmitter and receiver support 802.11b, 802.11g and 802.11n standards.

They find that the RSSI values vary a lot with the presence of a brick wall. That is, the RSSIs obtained in the presence of a brick wall have a significantly lower mean and higher variance when compared to the RSSIs obtained without the brick wall. They also find that as a human subject crosses between the transmitter and the receiver, the RSSI values drop significantly (by about 20 dBm). These two studies [22, 39] highlight the impact that changes in the environment can have on the RSSI values.

In the studies described above [1, 13, 22, 38, 39, 46] RSSI values are used to study different aspects of wireless networks. From their results we understand that the RSSI changes with distance [38, 46] and also with changes in the environment [1, 13, 22, 39]. This suggests that by analyzing the RSSI and its variability we can characterize the environments in which the networks operate. In our analysis, to account for variability in the RSSI values we define and use dispersion thresholds. This helps analyze the variability in signal strengths and to characterize the environments of signal propagation in which the commercial Google Wifi devices are deployed.

The results from Tauber et al. [46] also describe the impact of low RSSIs on throughput. This is because, the clients whose signals are received with low RSSIs are limited in the range of PHY rates they can use for transmission. This range of rates depends on the receiver sensitivity of the wireless chipset for different combinations of the spectrums, the channel widths and the MCSIs. Therefore, when analyzing the RSSI values, it is important consider the impact the spectrums, the channel widths and the MCSIs can have on the interpretation of the signals. For example, for a signal received by the AP with a particular signal strength, the AP might be able to decode more rates if the signal uses the 2.4 GHz spectrum and 20 MHz channel width than if the signal were to use the 5 GHz spectrum and 80 MHz channel width. In this thesis, knowing the signal strengths, we use the receiver sensitivity (from the datasheet of the receiver's chipset) to determine which rates could be decoded. Therefore, obtaining results that are independent from the capabilities of the clients in the network.

2.2.2 Characterizations based on Throughput

In this section we review studies that use throughput as one of the metrics to study WiFi networks. While there are numerous studies that use throughput to analyze WiFi networks we primarily focus on those that analyze the relationship between throughput and RSSI. We also describe studies that analyze the impact of interference from other clients and neighboring WiFi links on throughput. In addition to RSSI and interference, the studies described below explore the different factors that can impact the throughput experienced by the devices in a WiFi network.

Pelechrinis et al. [36] in their 2010 paper, study whether or not the higher physical rates of 802.11n networks can translate to high quality links in a real deployment. We describe their paper because they analyze the impact of interference on the recorded Packet Delivery Ratio (PDR) for high RSSI values and high PHY rates.

They analyze the link quality using the PDR. The PDR measures packet losses on the 802.11n link. High PDR implies low packet losses, while low PDR implies high packet loses. They perform experiments on an indoor testbed using hardware with 802.11n NICs that support up to 2x3:2 MIMO configurations (i.e., 2 transmitting antennas, 3 receiving antennas and 2 streams) that can operate using both 2.4 GHz and 5 GHz spectrums. To investigate the link quality of 802.11n networks they analyze the PDR and the throughput with regards to the PHY rate, the received signal strength and the channel width. Further, they also disable the automatic rate adaptation scheme to test different PHY rates using different combinations of MCSIs, spectrums and channel widths. They first establish a performance baseline (in terms of PDR) by conducting controlled and interference free experiments using the 5 GHz spectrum. They do this to better understand the 802.11n PDR for subsequent experiments where links are subject to interference. To study the effects of channel width, transmission rate and TCP throughput they conduct experiments using the 2.4 GHz spectrum in the presence of interference from other co-located 802.11n networks.

They find that for PHY rates greater than 180 Mbps, the PDR decreases as the PHY rate increases even for high RSSI values (i.e., between -35 dBm and -50 dBm). This indicates that the decrease is not due to weak signal strengths but due to the inability of the space division multiplexing and high-rate modulation schemes to support error-free communication in the test environment. They also find that interference has a non-negligible impact on the high PHY rates (i.e., PHY rates greater than 150 Mbps). That is, in the presence of interference, clients that use PHY rates greater than 150 Mbps record a lower PDR when compared to experiments done without interference (i.e., the baseline experiments). To analyze if this decrease is due to channel bonding, they repeat their experiments using only 20 MHz channel widths. They find that in the absence of channel bonding PDR increases to about 0.95 in all cases. This shows that a wider channel is more susceptible to interference from external WiFi links. They further analyze the impact of a decrease in PDR on the TCP goodput. TCP goodput is the rate at which useful data traverses a link. They find that TCP goodput for high PHY rates using the 40 MHz channel width is lower than when the same configuration uses the 20 MHz channel width. They speculate that the low TCP goodput for high PHY rates can be due to two reasons. Either the PDR is very low and that translates to IP losses seen by TCP, or there are no IP losses but the overhead due to the binary exponential backoff of MAC retransmissions is high. Their work shows that interference from neighboring WiFi networks can have a significant impact on users' WiFi experience.

Fiehe et al. in their 2010 paper [14], study the throughput of clients operating in 802.11n networks by conducting multiple experiments in both interference-controlled and typical office environments. What interests us is that they also study whether the distance between the client and the AP and interference from external sources and the other clients in the environment can affect the throughput experienced by the clients.

The experimental set up uses a UDP connection between an AP and a computer equipped with a wireless card which acts as the receiver. Both devices support the 802.11n standard. To generate continuous UDP packets they use `iperf`. In order to obtain reliable per-packet information an additional laptop equipped with an AirPcap N capture card is placed near the receiver. In all measurements they use 40 MHz channel widths and two spatial streams. They conduct measurements in two different environments. First, to establish a baseline for UDP throughput with very little to no help from MIMO, they use an anechoic chamber. In other experiments, to generate controlled interference they use either a signal generator or two 802.11g equipped laptops which generate UDP traffic. To study the UDP throughput of an 802.11n network in an environment similar to expected deployments, they use an office in a university building with an AP placed at a distance of 7.5 m from the receiver with a soft partition wall between them. In order to limit the influence of interference from other devices using the 5 GHz spectrum all measurements in the office environment are done either at night or weekends. They also try to ensure that no active interferers are present.

They find that in an anechoic chamber that the the best achievable mean throughput is 161.61 Mbps with an almost complete absence of improvements due to MIMO. They find that in the presence of interference from 802.11g systems that transmitting using a 4 Mbps PHY rate the mean UDP throughput reduces to about half of the throughput compared to the ideal case of 161.61 Mbps. Furthermore, in the presence of interference created using a signal generator, they also find that given sufficient power even a relatively narrowband signal (with a bandwidth of 4 MHz) can reduce UDP throughput to almost zero. In an office environment they expected the throughput to improve due to gains from MIMO. They find that in an office environment the average throughput is about 130 Mbps. This is less than the ideal case (i.e., 161.62 Mbps in anechoic chamber) and is because of the soft partition wall between the sender and the receiver. Moreover, in the presence of an 802.11g network that uses 4 Mbps and 16 Mbps PHY rates, throughput reduces to about 90 Mbps and 75 Mbps, respectively.

While they do not measure RSSI values they find that interference from neighboring WiFi links operating using the 802.11g standard and obstacles in the environment of signal propagation can severely reduce throughput. Therefore, the environments that a network operates in can affect the throughput experienced by the clients. One of the goals of this thesis to to characterize the environment of signal propagation using the RSSI measurements.

Kriara et al. in their 2013 paper [28], study how the different features of the 802.11n standard affect performance of a WiFi client using 5 different link qualities based on their average RSSI values. One of the metrics they use to evaluate performance is throughput experienced by the clients.

The testbed used in their experiments consists of 8 nodes of which 6 form an 802.11n WLAN with one access point and 5 clients placed in different locations to emulate different link qualities based on the RSSI values. The other 2 nodes are set up to act as another co-located 802.11n WLAN to realize different interference conditions. The 5 different clients each record a different mean RSSI value between -33 dBm and -81 dBm. This depends on the location of each client. For example, Client 1 (that is closest to the AP) records a mean RSSI value of -33 dBm, while Client 5 (that is furthest from the AP) records a mean RSSI value of -81 dBm. They define links with lower RSSI values as links having a poorer link quality. To quantify 802.11n performance they use three metrics: throughput, packet loss and fairness. Fairness is the ability of the MAC layer to equitably share the common channel among all contending hosts. To generate traffic flows they use `iperf`. The 802.11n features tested are frame aggregation (sending two or more data frames in a single transmission), MIMO, channel bonding and all available modulation and coding rates. To establish a baseline, they measure the RSSI and throughput of different links, one at a time in the absence of any interference. Finally, to study co-channel interference (CCI) and adjacent channel interference (ACI) they use the 2 additional nodes mentioned previously to create an interfering link belonging to a co-located 802.11n WLAN with a single client. Co-channel interference is when two links use the same channel (channel 149 in their case) and adjacent channel interference is when the two links use adjacent channels. To generate ACI, they assign adjacent channels to the link under test and the interfering link (channels 149 and 153, respectively when channel bonding is disabled and channels 149 and 157 otherwise). To capture the worst case CCI and ACI effects, they place the interfering link in close proximity to the AP (i.e., less than 3 meters away). This effectively subjects every link under test to fairly strong CCI and ACI interference. They use `iperf` to generate traffic flows.

They find that in the absence of interference (i.e., without using the 2 nodes that emulate neighboring WiFi links) the average throughput experienced by different clients is proportional to the RSSI values. That is, clients with higher average RSSI values experience higher average throughput compared to clients with lower average RSSI values. They also find that the impact of different 802.11n features on WiFi throughput depends on the scenario. For example, from experiments using the two nodes that generate interference they find that for poorer link qualities in the presence of CCI signals, a higher MCS index is able to achieve higher throughput when compared to using frame aggregation. However, in the presence of ACI, increasing the MCS index does not improve the throughput for poorer link qualities because transmission activity from the interfering link in the adjacent channel increases the noise floor and reduces the signal to noise ratio, suggesting that lower modulation and coding rates are more effective in improving the throughput. They also find that the benefit of using two streams for throughput improvement is limited to only very good quality links. However, the impact of two streams is reduced in presence of ACI. Similarly, frame aggregation also does not improve throughput of poorer link qualities in the presence of interference. They find that there exists potential interdependence among various 802.11n features (e.g., frame aggregation, channel bonding and space division multiplexing) that prevents them from being treated in isolation when the aim is to improve throughput. For example, in the presence of interference from neighbouring WiFi links, a client with an average RSSI value of -50 dBm is able to achieve higher throughput using frame aggregation than without it for MCS index 0 to 3. However, for MCS index 4 to 7 frame aggregation does not improve throughput while channel bonding and the use of multiple streams does improve throughput. They also find that in all scenarios, clients with low RSSI values are not able to use all the MCS index values. That is, while the client with better average RSSI values ($\geq$ -50 dBm) could use all MCS indices, clients with weaker RSSI values could not use all MCS indices. The client with an average RSSI of -81 dBm could only use the 2 lowest MCS index values. This reduces the maximum usable PHY rate and consequently the throughput of that client. Their work shows that the throughput in 802.11n networks depends on the link quality (as measured using the RSSI values) and on the interference from the neighboring links.

Patro et al. [35] in their 2013 paper, study the WiFi experience of users in 30 home networks over a period of 6 months by providing users with custom APs. While we use the RSSIs to study the data from 446 networks, they use estimated throughput as a metric to study different networks.

They provide users with custom APs with dual WiFi NICs, one of which provides the AP functionality to clients and the other is used to collect measurements. The NICs support both 802.11n and 802.11g standards. While 14 APs are deployed in individual homes of volunteers in one apartment building, 6 APs were deployed in another dormitory-style housing apartment where the wireless network was centrally managed. Additionally, 10 APs were distributed to volunteers and colleagues of the authors. To study the data collected, they develop a metric called Witt (WiFi-based TCP throughput), which passively estimates the likely TCP throughput that different clients can expect given the channel conditions. Witt is calculated using a linear model that estimates TCP throughput based on packet losses, airtime utilization, and channel contention. Witt shows a 1:1 correspondence between the estimated throughput and the actual recorded TCP throughput. We do not describe this in further detail as it is beyond the scope of this thesis. They use the estimated TCP throughput as a direct measure of the link's performance and categorize links based on Witt values into 5 different categories. The categories range from Very Good (Witt $\geq$ 16 Mbps) to Very Poor (Witt $\leq$ 1 Mbps). Using Witt the highest throughput they estimate is 19 Mbps for 802.11g networks and 34 Mbps for 802.11n networks. That is why they classify Witt $\geq$ 16 Mbps as Very Good. Their APs are equipped with Atheros 9220

Mini-PCI WiFi NICs, that supports the 802.11n standard in both 2.4 GHz and 5 GHz spectrums. These devices have 2 transmitting and 2 receiving antennas. They can therefore use PHY rates of up to 130 Mbps using 20 MHz channel widths and 300 Mbps using 40 MHz channel widths. Therefore, we believe that categorizing the TCP throughput that is greater than 16 Mbps as Very Good, is quite low for these devices as they should be able to achieve higher speeds.

From their analysis they find that about 20% of the clients in their deployment do not experience Good wireless performance (Witt is $<$ 8 Mbps) for about 80% of their active periods (i.e., the device has sent at least 500 packets in the last 10 second window). Additionally about 8% of the clients experience Poor performance (i.e., Witt is $<$ 4 Mbps) for more than 10% of the their active periods.

They further analyze the causes of Poor performances and find that in Building 1 low PHY rates ($\leq$ 12 Mbps), high airtime utilization ($>$ 60%) and high packet losses ($>$ 50%) cause Poor performance. In Building 2, the major cause of Poor performance is the use of low PHY rates. Building 1 has private APs per apartment resulting in a dense wireless deployment. Thus, some APs experience occasional high airtime utilization ($>$ 60%) due to neighboring WiFi links. Building 2 provides centralized wireless service to its residents and thus, some users can occasionally experience poor signal quality based on the client device and location. In both apartment buildings the use of low PHY rates is one of the primary causes of Poor performance. The presence of transmitters using low PHY rates has a high impact (i.e., Witt reduces from 12.5 Mbps to 6 Mbps) on wireless traffic due to high air time contention. This is because low PHY rates ($\leq$ 12 Mbps) require longer transmission time.

Grover et al. [21] in their 2013 paper, study the different characteristics of home networks. We are interested in the results of their analysis of network infrastructures. They define network infrastructure as the networking technologies (e.g., spectrums) and devices used in home networks (e.g., smartphones).

To study the characteristics of home networks the authors collect data by deploying 126 routers across 19 countries over 3 years. They recruit most of their AP users by word-of-mouth, or through targeted advertisements for specific experiments and projects that they run as part of their research. To study network infrastructures they develop and deploy BISmark [44], a custom router that supports 802.11a, 802.11g and 802.11n standards. They collect size and timestamp information for every packet relayed to and from the Internet. They also collect DNS responses, MAC addresses and IP addresses for their analysis.

They find from their analysis of the network infrastructure that the 2.4 GHz spectrum is significantly more crowded than the 5 GHz spectrum. The median number of devices seen on the 2.4 GHz spectrum is five, whereas on the 5 GHz spectrum, the median number of devices is two. Interestingly, in our data set 54.9% of the datapoints are from signals using the 5 GHz spectrum compared to 45.1% of the datapoints from signals that use the 2.4 GHz spectrum. Our findings indicate that there is more penetration of 5 GHz devices since that time.

Zeng et al. [54], in their 2014 paper, empirically study the throughput and power consumption of clients using 802.11ac networks. While their primary goal is to study the power consumption of clients under different settings, we describe the paper because they also show how different client configurations can achieve the same throughput.

They build a testbed using an AP that is compatible with the 802.11ac standard and supports 80 MHz channel widths, modulation schemes up to 256 QAM (Quadrature Amplitude Modulation) and 3x3:3 MIMO (i.e., it has 3 transmitting and 3 receiving antennas and 3 streams). They use 3 different devices as clients each with a different 802.11ac chipset. Two of these three clients support 3 streams and one client supports only 1 stream. They do this to avoid profiling specific hardware. They conduct experiments outdoors in a parking lot with a line of sight link up to 100 meters and indoors at night to avoid any interference. They use `iperf` to generate packet flows and measure throughput. All experiments are repeated 10 times and they report the average values of the measurements.

From the results of the experiments conducted in the parking lot, they find that as the distance between the AP and the client increases the client can use only lower MCS indices. For example, MCS 8 and 9 yield no throughput beyond 10 meters and MCS 4 and higher yield no throughput beyond 40 meters even in line of sight and zero interference environments. This is because, higher MCS indices require a higher signal strength but as the distance between the AP and client increases the signal strength decreases. In this thesis we use the receiver sensitivity to divide the range of MCS indices that can be decoded for a given range of signal strengths. We then use this to categorize and study both the clients and the networks.

They also find that there exists different combinations of MCSIs and spatial streams that can yield the same throughput. Therefore, the throughput, to a large extent, depends on the client capabilities. For example, they find that a client that supports a maximum of 3 streams and is at a distance of 10 meters from the AP, but uses only 2 of the 3 streams and MCS index 8 achieves the same throughput as when using all 3 streams and MCS index 5. Our study is agnostic of client capabilities.

Sundaresan et al. in their 2015 paper [45], study the relationship between PHY rates, retransmission rates, RSSIs and TCP throughput. This is done by deploying a passive measurement tool on commodity access points (which support 802.11g and 802.11n standards). Their study includes 66 homes across 15 countries and spans one month in 2013. We are interested in their study of the relationship between the RSSI and throughput and also in the results of their findings regarding the difference in PHY rates between clients using the 2.4 GHz and 5 GHz spectrums and the fraction of networks with one or more poor clients.

They collect data using BISmark [44], a custom router that supports 802.11a, 802.11g and 802.11n standards and can operate using both 2.4 GHz and 5 GHz spectrums. They collect packet traces (using `pcap`) of connections from the WAN interface (i.e., the access link) and both wireless interfaces in the device (i.e., from the 2.4 GHz and the 5 GHz spectrums). Packet traces from the WAN interface provide information about wide-area TCP connections and IP packets that traverse the access point. They configure the wireless interfaces to capture the source and destination MAC addresses, the received signal strength indication (RSSI) and information regarding the PHY rate used and whether a frame was retransmitted or not. The access points collect data every five minutes on average for 15 seconds per iteration. The AP also records the access link throughput every 2 hours. The access point runs `tcptrace`, which processes the `pcap` traces to measure the TCP throughput. They also analyze the correlation between the RSSI and the normalized throughput. The normalized throughput is the sum of TCP throughput for all flows in a home network divided by the access link throughput. This measures what fraction of the available access link throughput clients actually use.

They find that the average TCP throughput does not correlate with RSSIs (i.e., the correlation coefficient is 0.06). However, when they analyze a subset of flows (AP to client), whose normalized throughput is greater than 0.1 (i.e., for which they determine there is sufficient user demand), normalized throughput is positively correlated with RSSI (i.e., the correlation coefficient is as high as 0.8). This is interesting because we expect that such correlations are mainly of interest when there is enough traffic for the correlation to have a significant or noticeable impact on performance. The authors also find that the clients using the 5 GHz spectrum have higher normalized throughput than those using the 2.4 GHz spectrum. They find that only 30% of 2.4 GHz clients see median PHY rates above 65 Mbps. In contrast, more than 50% of the clients that use the 5 GHz spectrum see median PHY rates greater than 100 Mbps. Therefore, they conclude that clients using the 5 GHz spectrum perform better than clients using the 2.4 GHz spectrum. While they do not have any information regarding client capabilities, they find that about 60% of clients using the 2.4 GHz spectrum use PHY rates close to 65 Mbps. They speculate that this might be because these clients are small mobile devices with single antennas and 65 Mbps is the maximum PHY rate supported by the 802.11n standard for that configuration. While they analyze their clients according to the PHY rates they use, we analyze our clients using RSSI values and their variability. Our analysis is agnostic to client capabilities.

To study how wireless performance varies across devices in a single home, they measure the K-S distance (using the Kolmogorov–Smirnov test) [18] of the distributions of all of the PHY rates used between each pair of devices in each home. The K-S distance is defined to be the largest absolute difference between two empirical CDFs evaluated at any point. The K-S distance is always between 0 and 1. It measures how similar two data sets are (a K-S distance of 0 implies that two the data sets are identical). They find from their analysis that that more than 80% of homes have at least one pair of devices with a K-S distance of more than 0.6, indicating that most homes have at least one client that uses PHY rates that are very different from the rest of the devices in the same home. That is in 80% of homes there is at least one poorly performing device. Similarly, they also analyze the K-S distance of RSSIs across different AP-client pairs and observe similar differences (i.e., about 80% of AP-client pairs have a K-S distance of more than 0.4). Unfortunately, their analysis is not able to determine if the differences in PHY rates are due to differences in client capabilities or signal strengths. Because our approach considers signal strengths, it is agnostic to client capabilities and we are able to characterize the environments in which devices operate.

In contrast to the above body of work [14, 21, 28, 35, 36, 45, 54], our analysis does not depend on observed throughput, instead we analyze the RSSIs of the signals received by the APs. An advantage of analyzing RSSIs rather than observed throughput, is that our analysis is independent of the particular PHY transmission rates and number of spatial streams supported by the client. For instance, if a client only supports 1 spatial stream, that limits the maximum throughput of that client. Likewise, if the client does not support 80 MHz channel widths in the 5 GHz spectrum, then its throughput will potentially be lower than those that do support 80 MHz channel widths. In addition, analyzing the RSSIs gives us an insight into the nature of environment of signal propagation (as discussed in Section 2.2.1), since the signal strength is affected by various factors in the environment.

Previous studies that analyze WiFi networks do so in a limited setting, that is either by actively recruiting volunteers and deploying APs in specific buildings [35] or by deploying custom routers [21, 45]. In doing so, the results those studies report depend very much on the selection of homes the APs are deployed in. A downside of such studies is that it is unclear how representative the networks are. Especially those studies where many of the users are research scientist and their friends [35]. In contrast, the data analyzed in this thesis is collected from a large number of real deployments using commercial Google Wifi APs.

2.2.3 Categorization Techniques

The focus of this thesis is to characterize the signal strength of 2,946 clients operating in 446 networks. To effectively characterize the signal strengths we want to be able to categorize the clients and the networks, based on the strengths of the signals from clients, into different categories. Therefore, in this section we describe some of the previous studies that categorize signal strengths using different categories to analyze wireless networks. In describing these studies, our primary focus is on understanding the methodology of categorization that they use to categorize the signal strengths

Eckhardt and Steenkiste in their 1996 paper [12], investigate the the bit error rate of a WaveLAN network. This was a commercial product in late 1990s, that was designed for constructing 2 Mbps in-building wireless networks. The authors study whether they would be able to extend services like support for real-time or near-real-time guarantees in a wireless environment. WaveLAN provided an important foundation for the formation of the IEEE 802.11 working group and the resulting creation of WiFi standards. Furthermore, while the authors do not categorize the signal strengths, we describe the paper here because the results of their analysis of the signal strengths and the bit error rates are used in other studies that categorize signal strengths [1, 29, 41].

To investigate the bit error rate (BER) of the WaveLAN network, they monitor wireless data transfers between two identical laptops in the 902 – 928 MHz spectrum. BER is the number of bit errors divided by the total number of transferred bits. They also modify the device driver of the receiver to collect information regarding the status of every bit and the signal strength. The signal strength is reported by the receiver using a 6 bit integer. They send bursts of packets at a transmission rate of about 1.4 Mbps. They analyze the BER of data transfers using varying signal strengths. To vary the signal strengths, the receiver is held against one wall of a large lecture hall while the transmitter is moved away from it to various distances.

They find from their experiments that as the distance between the transmitter and the receiver increases, the signal strength decreases. Moreover, as the signal strength decreases the bit error rate increases. They also find that a signal strength that exceeds a certain threshold is sufficient for the receiver to receive packets reliably with extremely low bit error rates. However, when the signal strengths drop below the threshold, signals can no longer be reliably received due to high bit error rates. Eckhardt and Steenkiste [12] find that lower signal strengths have higher bit error rates. Other studies [1, 29, 41] use this knowledge to categorize WiFi signals on the basis of their signal strengths (i.e., the RSSI values).

Kuang and Williamson in their 2002 paper [29], study the performance of multimedia streaming applications for mobile Internet users in a wireless local area network (WLAN) environment. Their primary focus is to study the audio and video streaming quality in an 802.11b network under different channel conditions based on the RSSI values of the client. They do so by categorize the signal strengths into four qualitative categories: Poor, Fair, Good and Excellent.

Their experimental set up consists of one server, two laptops, and one AP. The server runs the RealServer software and acts as the video server. This video server is connected to a residential gateway using a 10 Mbps Ethernet card. One of the two laptops acts as the client device and it runs the RealPlayer client software. The second laptop runs the Sniffer Pro network analyzer software. Both laptops use a Cisco Aironet 350 network adapter. They classify the signal strengths into four categories: Poor (signal strength $< 20\%$), Fair (signal strength 20% to 45%), Good (signal strength 45% to 75%) and Excellent (signal strength $> 75\%$). These signal strength categories are based on the Link Status Meter on the Cisco Aironet 350 device which is used in their laptop that acts as the client. The paper does not provide enough information to convert the signal strength values reported in percentage to the corresponding signal strength values in dBm. To establish a reference point for their measurements, they first determine the maximum end-to-end throughput achievable in their experimental environment. For this purpose, they use `netperf` to invoke a 60 second TCP throughput test between the client and server. Each test is done 8 times. To test the multimedia streaming quality for different channel conditions they use an audio and video stream that are 68 seconds long each and have 432 and 2,850 total packets respectively.

They find that the weaker the signal strength, the lower the throughput and also the greater the variability in the measured throughput. They also observe that the maximum observed throughput is 4.6 Mbps for a channel that is rated as Excellent. This result indicates that the 10 Mbps Ethernet connection between the server and the AP is not a bottleneck in their experiments. They find from their analysis of multimedia streaming that the playback of the video and audio streams are very smooth for the Excellent and Good channel conditions. That is, they do not record any packet loss. For the Fair channel conditions, the playback of the video is jerky, indicating lost video frames, though the audio and visual quality of displayed video frames is good. Under the Poor channel conditions, they observe that the video playback is jerky, some individual pictures are blurry or truncated, and the audio quality deteriorates. In some cases, even the attempt to set up the streaming connection fails.

Sarkar and Sowerby et al. in their 2009 paper [41], study the impact of wireless channel conditions (based on RSSI values) on multimedia streaming quality. We are interested in their categorization of the wireless channel conditions, based on different ranges of RSSI values.

They conduct experiments using an AP and clients that support the 802.11b standard. Each experiment is repeated three times and they report the mean of those measurements. In all experiments by Sarkar and Sowerby, the transmitter is located in an office and the receiver is moved to various locations within the office building to obtain different average RSSI values ranging from -44 dBm to -91 dBm. The transmitter is configured to share the audio and video files with the receiver. They measure the audio and video quality using playback delays. This is the difference between the playback time of the media file measured at a client device and the original length of the media file. To investigate the impact of wireless channel conditions on the streamed audio and video quality, they classify the wireless channel state, based on average RSSI ranges, into five categories: Excellent ($\geq$ -60 dBm), Good (-61 to -75 dBm), Fair (-76 to -80 dBm), Bad (-81 to -89 dBm) and Very Bad ($\leq$ -90 dBm). These signal strength categories are based on the link status meter on the D-link's wireless card.

They find that the playback of the audio and video streams are very smooth for Excellent and Good signal strengths. For Fair signal strengths, the sound quality of the audio is good, however the video playback is jerky. For Bad signal strengths, the audio quality deteriorates and the video playback is jerky and some individual pictures are blurry or truncated due to significant frame losses. In the case of Very Bad signal strengths, they find that their attempts to set up the streaming connection fails due to high frame losses. Moreover, they observe that the video playback delays are slightly higher than those of audio playback delays, especially for weak RSSIs (i.e., RSSI is below -81 dBm). This increase in video playback delays is because of the high retransmission of video frames when the RSSI is very weak. Furthermore, the audio and video streaming connection fails for RSSI value of -91 dBm for all three network scenarios. They also do not analyze experiments with RSSI values less than -91 dBm.

Ding et al. in their 2013 paper [11], report on a measurement and modeling study of the impact of poor WiFi signal strength on smartphone energy consumption. We describe the paper here because their study involves classifying signal strengths into two categories, namely, Good and Poor.

They collect and analyze WiFi signal strengths and trace traffic volume using a free Android app using 3,785 smartphones that are geographically distributed over 145 countries. Each trace has an average of 4.2 months of data. They collect the RSSI data from each client device every 1 minute. Since the Android app is installed on different clients supporting different configurations, information such as which 802.11 standard is used is not available. For WiFi networks, they consider RSSIs greater than -80 dBm as Good and those below that as Poor. Their categories are based on the significant increase in data transfer time and energy drain observed below -80 dBm signal strength. This is measured using a controlled experimental set up. For this, they use three different smartphones and an AP. All four devices support the 802.11g and 802.11b standards. To measure the power consumption and the data transfer times for different RSSI values, the smartphones download a 100 KB file from a server that is connected to the AP using a 100 Mbps LAN. The data download is implemented using a simple client and server program running on the phone and the server. It also measures the data transfer delay. They control the RSSI values by adjusting the distance between the smartphones and the AP. They repeat all their experiments 50 times and report the average over all those recorded values.

They find that the data transfer time only changes slightly for RSSI values between -50 and -80 dBm (i.e., it is always about 200 ms). We believe that using the data transfer time that is measured by downloading a 100 KB file as a performance metric may not be representative of real world scenarios. Because, the file size is so small, much of the time may be spent on connection overhead (e.g., establishing and tearing-down the connection). However, they report a 142.2% increase in the data transfer time for an average RSSI value of -85 dBm, and a significant increase of 1345.5% at -90 dBm, when compared to -50 dBm. Similarly, the energy consumption of downloading increases by 113.3% and 810.5% at -85 dBm and -90 dBm, respectively. This is because, for RSSI values between -50 and -70 dBm, most frames are transmitted using a PHY rate of 54 Mbps, which is the highest rate supported by 802.11g. When the signal strength drops below -70 dBm, frames are transmitted at much lower PHY rates due to rate adaptation. For example, using an RSSI of -85 dBm 74.0% of the frames are transmitted using PHY rates equal to or lower than 11 Mbps, while at -90 dBm 92.6% of the frames are transmitted using the lowest PHY rate of 1 Mbps. This increases the data transfer time and consequently the power consumption. Therefore, for the analysis of their data set they classify WiFi signal strengths of less than -80 dBm as Poor and those above it as Good.

They also find that all 3,785 users on average experience Poor WiFi signals (i.e., below -80 dBm) 25% of the time, and over 80% of them experience Poor WiFi signal strength over 5% of the time. We find from our analysis of the RSSIs that for both 2.4 GHz and 5 GHz spectrums only about 10% of signals using the 20 MHz channel widths are received with RSSIs less than -80 dBm and for signals using the 40 MHz and 80 MHz channel widths about 20% of the signals are received with RSSIs less than -80 dBm. Furthermore, they also find that foreground data traffic is much more prevalent than background data traffic, with the users transferring on average 96% of their total data volume during active device usage. They define active device usage as when the device has its display turned on. On average 21% of these data transfers occur during Poor WiFi signal strengths, with over 80% of the 3,785 users transferring over 2% of their foreground data during Poor signal strengths.

In this thesis, we analyze the clients and the networks on the basis of the strengths of the signals received by the AP. To do this, we are interested in categorizing the signals into different categories based on their RSSI values. The studies summarized above [1, 11, 29, 41] use ranges of RSSI values for the purpose of categorization. While Sarkar et al. [41] and Kuang et al. [29] base their categories on the wireless adapter used by them (i.e., using Cisco Aironet 350 and D-Link DWL-G132 network adapters respectively), Ding et al. [11] define the RSSI categories from the results of their empirical measurements. Unlike the other studies that we describe above [29, 41], Ding et al. are not able to categorize the RSSI values based on the chipset used because the RSSI values collected are from many different devices using many different chipsets. While there is no one size fits all rule for categorizing the RSSI values into different ranges, we can use the results of Eckhardt and Steenkiste [12] as a framework. They find that when signal strengths drop below a certain threshold packets can no longer be reliably received. Moreover, they also find that lower signal strengths result in higher bit error rates. In this thesis, for different combinations of MCSIs, spectrums and channel widths, we use the receiver sensitivity of the chipset (Qualcomm's IPQ 4019 in our case) as a threshold of signal strength below which the signal received by the AP cannot be successfully decoded. Thus, we are able to ascertain the different ranges of PHY rates that can be decoded by the receiver for the particular signal strength which is independent of client capabilities.

Furthermore, we also observe that signal strengths are highly variable. While there are clients whose signal strengths recorded by the AP are quite stable with very little variability, there are also clients whose signals are received by the AP with very high variability. This is because, as discussed in Section 2.2.1, changes in the environment of signal propagation can impact signal strengths. All studies that we describe in this section do not adequately address the issue of signal strength variability over time. In contrast, in this thesis we design a methodology that includes the ability to analyze the variability in the signal strengths. Since, signal strengths are impacted by the changes in the environment, by analyzing the variability in signal strengths from clients to the AP, we are able to better characterize and understand the environments in which the clients operate.

Chapter 3

WiFi Dataset

3.1 Introduction

In this thesis we analyze data collected from 446 Google Wifi access points with 2,946 clients over a period of 24 hours. The data has been provided by Google and was obtained from randomly sampled access points. As a result, we have no information about the regions in which the devices are used or how they have been deployed (e.g., whether or not the APs are or are not part of a mesh network). However, we surmise that the majority of the devices are deployed in homes with some likely being used in small office and other environments. All data is anonymized and cannot be traced back to any particular access point, client device, or user. Furthermore, no payload data or IP addresses are collected or examined. The recorded data includes information important to our analysis like the RSSI (Received Signal Strength Indication), the time in epoch seconds (using GMT), the hashed access point device ID, the hashed client ID, the frequency spectrum, and the channel bandwidth. The dataset has 417,122 datapoints from samples taken at each access point (AP) every five minutes. 45.1% of the datapoints are from signals using the 2.4 GHz spectrum and the remaining 54.9% of them are from signals using the 5 GHz spectrum, as shown in Table 3.1. It is also interesting to note that the majority of 5 GHz signals use the 20 MHz channel width. We suspect that many of these are ACKs.

Frequency	**Channel Width**	**Datapoints**	**Percentage**
2.4 GHz	20 MHz	188,076	45.1%
	40 MHz	0	0.0%
5 GHz	20 MHz	149,842	35.9%
	40 MHz	34,308	8.2%
	80 MHz	44,896	10.7%
Total		417,122	100.0%

Table 3.1: Distribution of data points

3.2 Hardware

The modern commercial Google Wifi devices use Qualcomm's IPQ 4019 system on chip which is capable of simultaneously supporting dual band 2.4 GHz and 5 GHz communication. It has 512 MB of RAM, 4 GB of flash memory, and four CPUs clocked at 710 MHz each. It also supports 2 streams (2x2:2) and beamforming. A companion mobile application provides the ability to configure and administer the network.

3.3 Analysis of Client Distributions

We now analyze the distribution of the number of clients connected to each access point (AP). This gives us an overview regarding the size of networks examined in our dataset. The size of a network is determined by the number of clients connected to a particular access point (AP).

Figure 3.1 shows histograms and CDFs of the number of clients connected to the APs from three different perspectives. They show the total number of unique clients connected to the AP throughout the 24 hour period, the maximum number of clients that are simultaneously connected to the AP across all 5 minute intervals, and the average number of clients that are simultaneously connected to the AP across all 5 minute intervals. For each access point (AP_i), where the number of clients that are simultaneously connected in each of the w 5 minute intervals are n_1,. . . , n_w, these values (average and maximum number of clients that are simultaneously connected) can be represented mathematically as shown in Equation 3.1 and Equation 3.2.

$$Max(AP_i) = max\{n_1,\ \ldots,\ n_w\} \tag{3.1}$$

$$Avg(AP_i) = \frac{\sum\limits_{j=1}^{w} n_j}{w} \tag{3.2}$$

In Figure 3.1 the x-axis shows the number of clients that are connected to the AP, the y-axis on the left shows the percentage of networks for the histograms and the y-axis on the right shows the fraction of networks for the CDFs. We now highlight a few observations from the Figure 3.1.

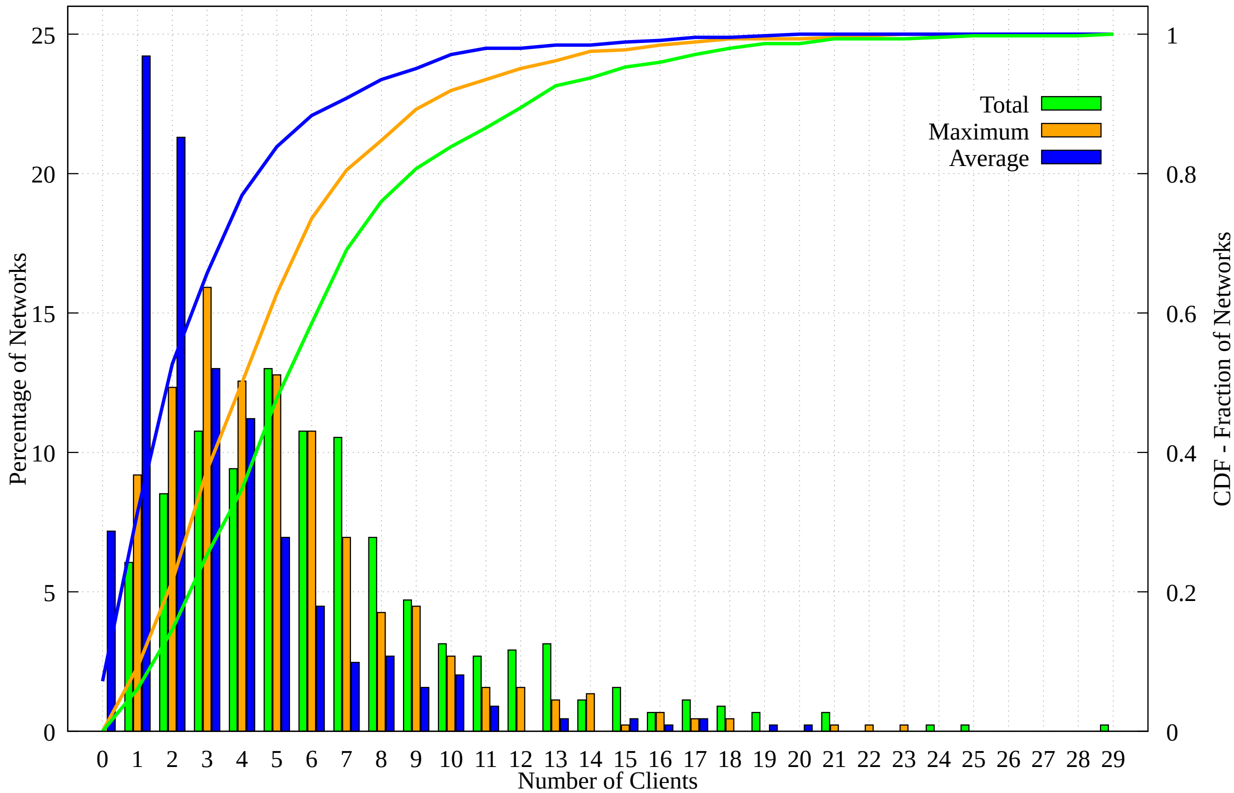


Figure 3.1: Information about connected clients

First, we note that across all access points (APs), the average number of simultaneously connected clients across all 5 minute windows is quite small. That is, on average 65.7% of the APs have 3 or fewer clients that are simultaneously connected in any 5 minute window, and 45.5% of the APs have only an average of either 1 or 2 simultaneously connected clients.

Second, we note that 6.5% of the access points (APs) have on average 9 or more clients that are simultaneously connected across all 5 minute windows. Combined with the first point above, this shows that while on average most APs service only a few clients, there are some APs that service a large number of clients (i.e., 9 or more). We also note that, at their peaks, 6.5% of the APs service a maximum of 12 or more clients simultaneously

Finally, we observe that all three CDFs have long tails. There are 6 of 446 networks that have a total of 21 or more connected clients, with one network having as many as 29 connected clients. These APs also have a relatively large number of maximum and average number of clients that are simultaneously connected. For instance, the AP that has a total of 29 connected clients, has on average 19 simultaneously connected clients, and a maximum of 23 simultaneously connected clients. Similarly, the AP that has a total of 24 connected clients has on average 20 and a maximum of 22 clients that are simultaneously connected. Later in Chapter 6 we investigate the impact, the size of a network has on the network.

3.4 Analysis of the Number of Datapoints per Client

Figure 3.2 shows the CDF of the number of datapoints collected for every client. Because data is collected every five minutes by the access points, 288 datapoints represents 24 hours worth of data. We note that in Figure 3.2, 3.6% of the clients have only one data point and 27.4% of the clients have 30 or fewer data points (i.e., 2.5 hours worth of data). In addition, 31.1% of the clients have almost a day's worth of data (i.e., more than 22 hours). The reason we do not have more clients with closer to a full day's worth of data may be because the number of mobile clients is significant and they connect to the network only when the user is home. Alternatively, there may be other clients (e.g., televisions and gaming consoles) that do not connect to the network when powered off or in sleep mode. Moreover, it is also possible that the AP might drop some data before uploading it to the server because of problems connecting to the internet through the ISP while the data was still cached. We also note there will be no readings obtained if the client is not powered on (e.g., a TV is turned off). There will also be no readings if the client moves out of the range of the AP and is disconnected. This may happen in scenarios when someone who uses a mobile client, like a cellphone, tablet, or smartwatch, leaves the house or building with the AP.

Another reason that we do not see more clients with closer to 24 hours of data is that some APs may be part of a mesh network (Google Wifi APs are designed to work as a part of a mesh network and are sold individually or in packages of 3). If a client belongs to a mesh network it may associate with one AP for some time and then due to a reduction in signal strength it may change its association to a different AP in the mesh. We have determined that no clients associate with more than one AP for which we have data. This means that the dataset contains no evidence that any of the APs are part of the same mesh network. On the other hand, we have no information that can be used to determine if the APs that we have data for are being used as a part of a mesh network or not.

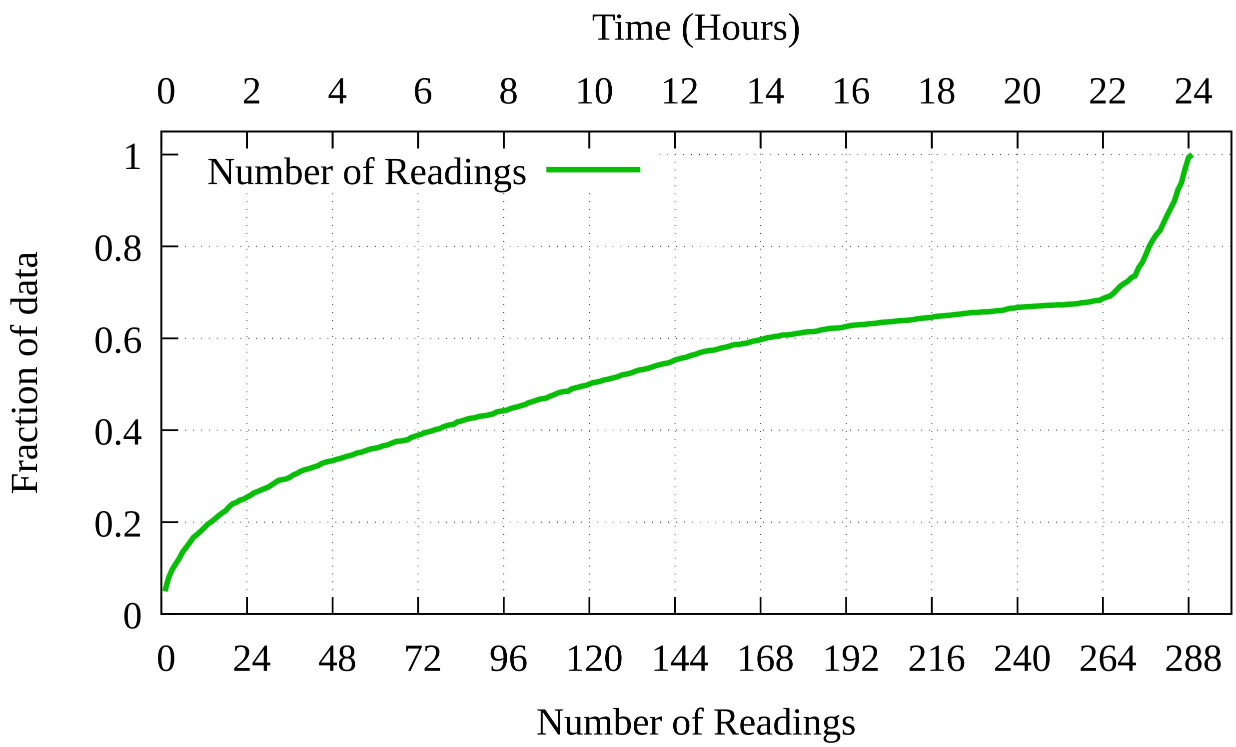


Figure 3.2: CDF of the number of datapoints per client

3.5 Analyzing RSSIs using Spectrums and Channel Widths

In Chapter 2 we explain that both the RSSI and the receiver sensitivity depend on the frequency spectrum and the channel width. Thus, we also analyze the distribution of the RSSIs of messages from all the clients received by the APs, based on their frequencies and channel widths.

Figure 3.3 shows the CDFs of the RSSIs using the 2.4 GHz and 5 GHz spectrums. We observe that signals using 20 MHz channel widths are received with a better RSSI, on both 2.4 GHz and 5 GHz spectrums, than the signals using either 40 MHz channel widths or 80 MHz channel widths on the 5 GHz spectrum. Also, signals using 40 MHz channel widths are received with better RSSIs than the signals using 80 MHz channel widths on the 5 GHz spectrum. This difference is because signals transmitted with same power on wider channels spread that power across more of the spectrum and hence the transmission power per hertz is lower than when transmitted using narrower channels. As also discussed in Section 2.1 increasing the channel width increases both, the sensitivity of the signal to interference, and the noise floor. Additionally, doubling the frequency, from 2.4 GHz to 5 GHz, quadruples the rate at which the signal attenuates, thus further deteriorating the signal strength. When devising a methodology to understand client signal strengths we need to account for these factors (as will be seen in Section 4.1).

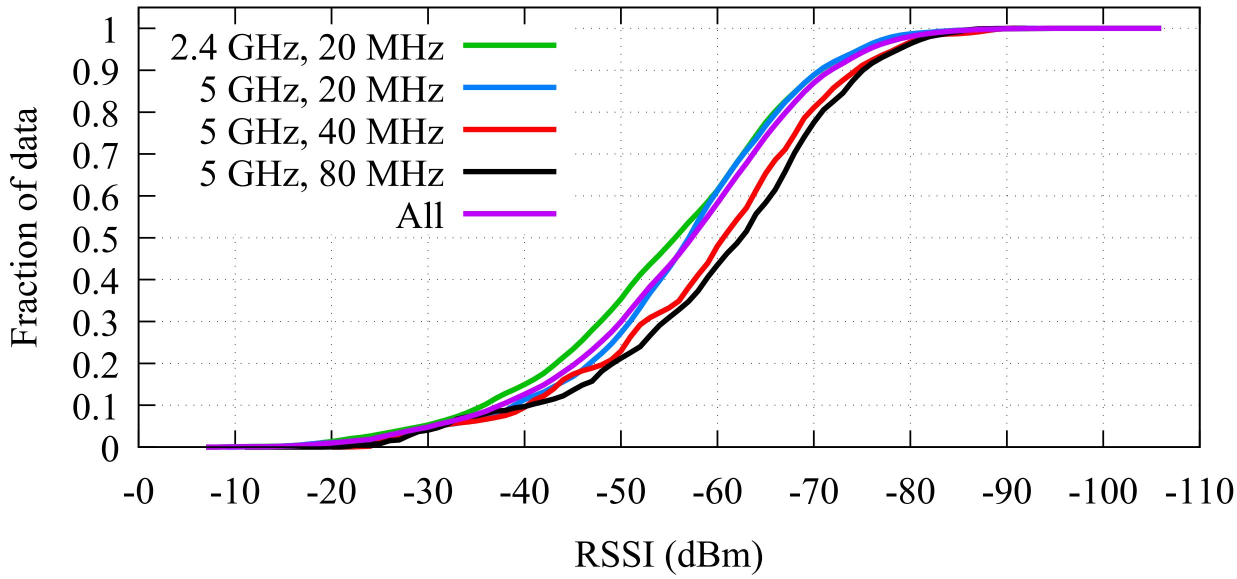


Figure 3.3: CDF of RSSIs when using different spectrums and channel widths

Chapter 4

Methodology

In this thesis, we study 446 networks and 2,946 clients based on the received signal strength indication (RSSI) of the messages received by the access point (AP) from the clients. As discussed in Section 2.1.10, only signals received by the AP whose RSSI is greater than the receiver sensitivity for a particular Modulation and Coding Scheme (MCS index), spectrum, and channel width can be decoded successfully. While a device might be able to transmit at high rates using a wider channel width, a denser modulation scheme and a higher coding rate (i.e., using a higher MCS index), if the signal is not received with a strength that is greater than the receiver sensitivity, then the signal cannot be decoded. In this chapter, we describe our methodology which is designed to categorize clients and networks based on the receiver's potential to decode messages received with different transmission rates based on the strengths of the received signals.

4.1 Challenges

Before we describe our methodology we first briefly explore the nature of our dataset and the challenges it presents when trying to categorize clients and networks. Figure 4.1 shows the signal strengths (RSSIs) of the messages received from all four clients of one access point (AP) from our dataset. The x-axis shows the 24 hour time period (using GMT) for which we have data and the y-axis shows the RSSI values of the messages received from the clients by the AP. Each data point on the graph represents the time at which a signal was received by the AP and the RSSI (i.e., signal strength) of that signal. Since a higher RSSI indicates a stronger signal (Section 2.1), data points that are higher on the graph indicate messages received with stronger signal strengths.

At a high level, we observe from Figure 4.1 that if we consider the signal strengths of the messages recorded by the AP from all its clients, then we can say that the signals are received by the AP with highly variable RSSIs. However, if we examine the clients individually, based on the RSSIs of its messages received by the AP, we see that the variability in strengths of the received signals differs from one client to another. For instance, messages from Client 1 are received by the access point with low variability as the RSSIs of more than 90% of the messages are in the range -65 dBm to -70 dBm. In contrast, the messages from Client 2 are received by the AP with more variability as the RSSIs are between -40 dBm and -60 dBm. However, Client 2 also has periods of low variability, as can be seen between 9:00 (GMT) and 11:00 (GMT), and also between 20:00 (GMT) and 23:00 (GMT). In contrast, messages from Client 3 are received with highly variable RSSIs. In our approach to categorization we intend to capture the general tendencies of the data while also including the notion of variability. We would also like to compare the signals from different clients to identify which have stronger signals. For example, a question we would like to study is, are signals typically stronger from Client 3 or Client 4, or are they similar?

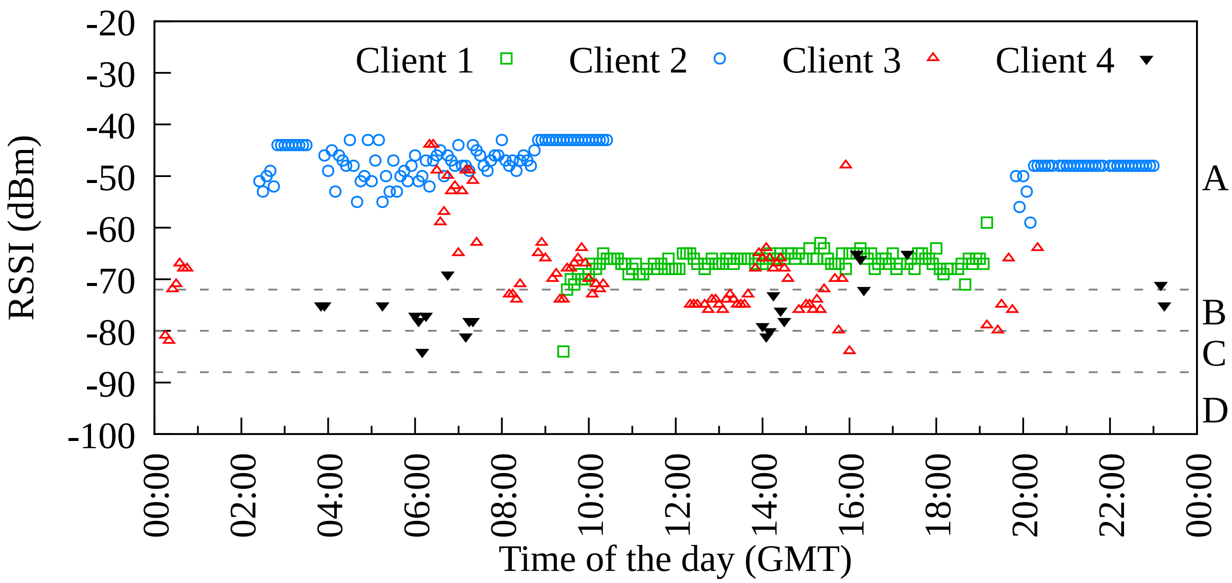


Figure 4.1: RSSIs of messages from clients received by an access point

We observe from Figure 4.1 that the majority of messages from Client 2 are received with a significantly higher signal strength (i.e., higher RSSIs) than all of the other clients. Thus, on average, it can potentially utilize higher PHY rates when sending messages to the AP than other clients. In contrast, messages from Client 3 are received by the AP with signal strengths in a larger range, with many messages received by the AP with relatively weaker RSSIs. For those weaker RSSIs, lower PHY rates must be used in order for the AP to decode the packet.

From our discussions of the RSSI in Section 2.1 we know that based on the receiver's sensitivity the signals received with a high RSSI can potentially use high PHY rates. However, the receiver sensitivity depends on the spectrum, channel width, and the MCS index used by the client for transmission. Therefore, our methodology must account for the receiver's sensitivity to factors like the MCS index, the spectrum, and the channel width used for transmission.

An advantage of analyzing RSSIs rather than observed throughput, as has been done in other studies [2] [21] [34] [35] [45], is that our analysis is independent of the particular PHY transmission rates and number of spatial streams supported by the client. For instance, if a client only supports 1 spatial stream, that limits the maximum throughput of that client. Likewise, if the client does not support 80 MHz channel widths in the 5 GHz spectrum, then its throughput will potentially be lower than those that do support 80 MHz channel widths. Because our methodology uses only the signal strengths of the messages received by the access point (AP), it is independent of the PHY rates supported by the clients. As a result, we are able to characterize the environments in which the devices operate. Previous studies that consider the client throughput are hampered by the constraints of the clients connected to the network.

The main goal and challenge in devising a methodology for categorizing clients is to capture the notion that Client 1 has the potential to use much higher transmission rates and has less variability than Client 3. Similarly, in categorizing the network we want to capture the notion that the network, shown in Figure 4.1, receives messages with RSSIs that have high variability. In addition, our goal is to distinguish between networks with clients whose messages are always or mostly received with good signal strengths, from the networks which have clients whose messages are received with either variable or weak signal strengths, since they will be limited to fewer and lower PHY rates. In other words, we would like to determine and distinguish clients and networks that are mostly operating with Good signal strength from those that are not.

4.2 Requirements of our Methodology

To help achieve our goals the requirements of our methodology are:

1. There should be a relatively small number of categories (we choose four).
2. The categories should be distinct and provide an intuitive notion of better signal strengths (and operating conditions) in one category when compared with another.
3. The methodology should enable us to capture the notion of, and examine the variability in signal strengths.
4. We should be able to use the methodology to categorize both clients and networks.

4.3 Description of our Methodology

In this section we describe the details of our methodology. We begin by summarizing the steps of the design of our methodology to provide readers with an overview of the design. We believe this will help readers to better understand the various steps and reasoning for the choices made in our design. Later, in subsequent sections we elaborate on the steps in further detail. After describing our methodology, we use an example to illustrate how we use the methodology to categorize both clients and networks.

4.3.1 Summary of the Design of the Methodology

In designing the methodology our primary concern is whether or not the signals received by the device are strong enough to decode messages sent using different rates (i.e., using different combinations of MCSIs, spectrums, and channel widths). This depends on the receiver sensitivity of the wireless chipset used in the device (Qualcomm's IPQ 4019 in our case). Next, for every signal received by the access point (AP), we examine the signal strength (RSSI) to find the range of rates that can be successfully decoded by the AP for that RSSI. Based on the range of rates that can be decoded we classify them using four levels, A through D. Signals in level A are strong enough to decode all rates, while signals in level D can decode only the lowest rates and are often unreliable as they are only a little stronger than the noise floor. Finally, depending on the number of levels spanned by the signals, we categorize the device (client or AP) into one of four distinct categories: Good, Moderate, Variable, and Weak.

4.3.2 Selecting the Levels

We begin by assigning every RSSI value a level. Levels for different RSSI values are determined by the receiver sensitivity for the signal received by the AP using a particular spectrum, channel width, and modulation and coding scheme. We create four levels A through D with A being the highest level. Levels are chosen based on the number of possible PHY rates that could be decoded given the value of the RSSI for a received packet. Signals in level D are only strong enough to decode the lowest two MCS indices for all spectrums and channel widths, which indicates that the signals are limited to using only the lowest rates. Since, the AP can only record RSSIs for messages it can decode, we do not have any information about signals that were too weak to be decoded by the AP. Level C includes RSSI values that permit the decoding of only rates using the lower 50% of the MCS indices. Level B includes RSSI values strong enough to decode all but the rates with the highest MCS indices. Finally, level A specifies the minimum RSSI value (signal strength) needed to decode the signal if it was transmitted using the highest MCS index for each spectrum and channel width, and thus indicates the potential for the client to transmit using the highest available rates. In other words signals in level A should be strong enough to decode all available rates. These different levels (A, B, C, and D) and their RSSI ranges based on the receiver sensitivity, from the datasheet for the IPQ 4019 chipset [23], are summarized in Table 4.1. This is the chipset used in the Google Wifi devices used to collect the data analyzed in this thesis. Note that our methodology is general enough that it could be used with other devices with different receiver sensitivities. This simply requires changing the RSSI values used for the levels A, B, C, and D, based on the receiver sensitivity indices for the chipset used in the particular AP.

Spectrums and Channel Widths		Levels and RSSI Ranges							
		Level D		Level C		Level B		Level A	
		RSSI (dBm)	MCSI	RSSI (dBm)	MCSI	RSSI (dBm)	MCSI	RSSI (dBm)	MCSI
2.4 GHz	**20 MHz**	<= -89	0 to 1	-81 to -88	0 to 3	-73 to -80	0 to 6	>= -72	0 to 7
	40 MHz	<= -86	0 to 1	-78 to -85	0 to 3	-71 to -77	0 to 6	>= -70	0 to 7
5 GHz	**20 MHz**	<= -87	0 to 1	-77 to -86	0 to 4	-69 to -76	0 to 7	>= -68	0 to 8
	40 MHz	<= -85	0 to 1	-74 to -84	0 to 4	-68 to -73	0 to 8	>= -67	0 to 9
	80 MHz	<= -81	0 to 1	-71 to -80	0 to 4	-62 to -70	0 to 8	>= -61	0 to 9

Table 4.1: RSSI ranges for various levels across different spectrums, channel widths, and MCSIs

4.3.3 Need for Categories

Figure 4.1 also denotes each of the four levels (on the the right side of the graph). For the purposes of describing the methodology, momentarily assume that every message was sent to the AP using the 2.4 GHz spectrum and 20 MHz channel width. At a high level, we see in Figure 4.1 that more than 95% of the messages from Client 1 and all messages from Client 2 were received by the AP with signal strengths in level A. Thus, both of these clients can almost always transmit using the highest possible rates and the AP can still potentially decode the transmissions. Intuitively, the AP receives signals with good strengths from these clients. In contrast, messages from Client 3 and Client 4 are received by the AP with signal strengths in different levels. Thus, unlike Client 1 and Client 2, we cannot say that the strength of signals received from Client 3 and Client 4 are either good or weak. This is because there are times when the messages sent by Client 3 and Client 4 are received with good signal strengths and other times when the messages sent by them are received with weak signal strengths. The transmitting clients' rate adaptation algorithm will infer from packet losses for high PHY rates that it must use low PHY rates. Thus, to capture this notion of variability and the range of possible rates that a client can use to transmit messages to the AP (or the range of levels that the RSSIs span), we create four categories.

4.3.4 Description of Categories

As discussed in Section 4.2 one of the requirements of our methodology is to categorize clients and networks into a relatively small number of categories that are both distinct and intuitive. Thus, we define four categories: Good, Moderate, Variable, and Weak. The categories have been chosen to appeal to intuition, with a sense of a natural order among them. Good is more desirable than Moderate, which is more desirable than Variable, which is more desirable than Weak. We note that with more desirable categories, it is more likely the AP will be able to more often decode messages sent using higher rates.

4.3.5 Steps to Categorize Clients and Networks

We now provide a high-level overview of how we use the methodology we have designed above to categorize clients and networks.

1. Using Table 4.1 we assign a level to every RSSI value.
2. Using a particular dispersion threshold, we determine the range of levels spanned by the threshold.
3. Using those ranges, we determine the category.

For convenience, the median value is referred to as determined using a dispersion threshold of 0, because it does not include any notion of the variability.

The analysis of clients and networks are very similar, and follow the same steps, but differ only in the data they consider for analysis. For analysing a client, we consider only the RSSI values of the messages received by the AP from a particular client. In contrast, when analysing a network, we consider the messages recorded by the AP from all of its clients.

4.3.6 Using the Methodology

For our data analysis, we borrow techniques from descriptive statistics to interpret the data. We begin by first analysing the data using the median signal strength (RSSI). An analysis using the median RSSI reveals the central tendency of the data and also condenses the dataset to one representative value. This is useful when working with large amounts of data. If the median signal strength lies in either level A or B we categorize the client or the network as Good, since it can decode the messages sent using high rates (high rates are encoded using high MCS indices). Similarly, if the median signal strength lies in either level C or level D we categorize the client or the network as Weak, since it can can only decode messages sent at low rates.

However, categorizing a client or a network using the median signal strength does not capture any information about the variability of the signal strengths. For instance, if the median signal strength of Client 3 is in level A, we would categorize it as Good when in fact it is variable in nature. Thus, to capture this variability, we also include some notion of dispersion of the RSSI values around the median. As a result, we examine the minimum and maximum RSSI values that are determined by including 50%, 60%, 70%, 80%, 90%, and 100% of the RSSI values located centrally around the median. We call these values, dispersion thresholds and calculate them using percentile ranges of 25–75, 20–80, 15–85, 10–90, 5–95, and 0–100, respectively. By examining different dispersion thresholds, we can determine whether the RSSIs are clustered together in one level or scattered across multiple levels. This is useful not only to capture the variability of the RSSIs in our categorizations, but it also allows us to study how the categorizations change with the inclusion or exclusion of varying degrees of what may be considered atypical RSSI values. As noted previously, we refer to our analysis using the median as using a 0% dispersion threshold, since it does not including any notion of variability.

Figure 4.2 visually depicts the levels (A, B, C, and D) and our categories (Good, Moderate, Variable, and Weak) for the data. We see that levels A through D are fine if we use a metric that does not include variability (e.g., the median or the mean) or if the variability is low enough that the range of dispersion measure fits into one level. However, in some cases, we find that the RSSI is highly variable (e.g., Figure 4.1). For instance, we see from Figure 4.1 that RSSIs of messages from Client 3 are highly variable when compared to Client 2. All but two RSSI values of Client 2 are in the range between -65 dBm and - 70 dBm (i.e., level A), whereas the RSSIs of Client 3 are in the range between -42 dBm and -80 dBm (i.e., between level A and level C). Thus, to capture this notion of variability we categorize a device (client or AP) based on the range of levels spanned by the RSSI values for different dispersion thresholds.

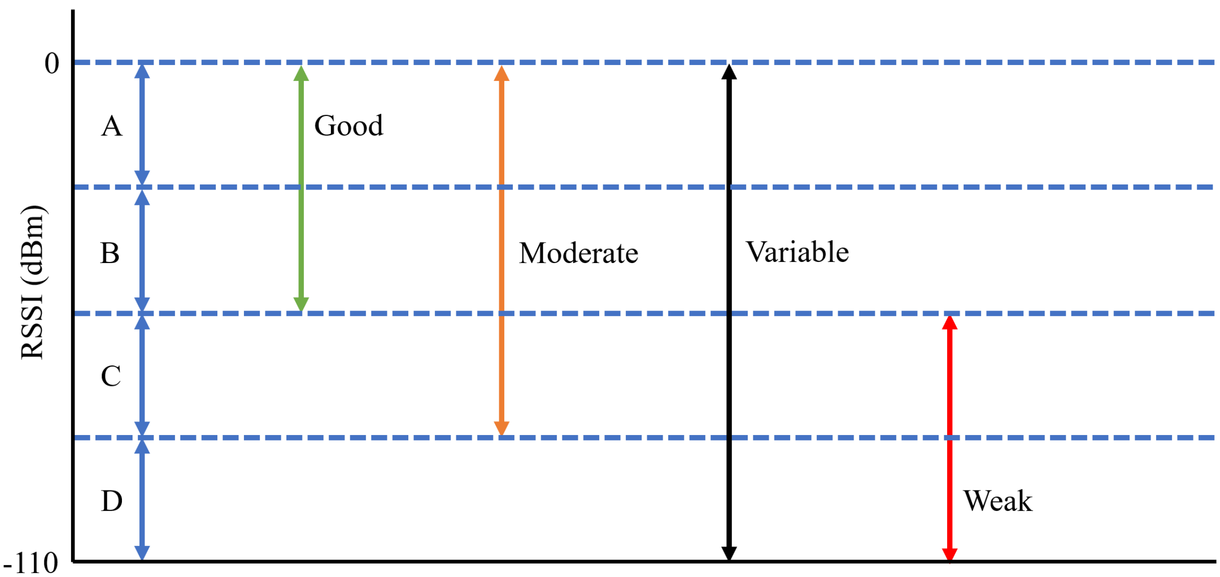


Figure 4.2: Levels and categorizations

4.3.7 Algorithm for Categorization

Based on our discussions above, we present the algorithm below to determine the category for a device (AP or client) using different dispersion thresholds. To find the category for a client, the algorithm considers only the RSSIs of the messages received by the AP from the specific client. In contrast, to find the category for a network the algorithm considers the RSSIs of the messages recorded by the access points from all its clients.

Algorithm 1 Algorithm to find categories for each device using different dispersion thresholds

for each device **do**
  deviceRSSIs[ ] $\leftarrow$ getDeviceRSSIs(device)
  // Get levels based on spectrum, channel width, and MCSI
  deviceLevels[ ] $\leftarrow$ getDeviceLevels(deviceRSSIs[ ])
  for each dt in $\{0, 50, 60, 70, 80, 90, 100\}$ **do**
    levelsSpanned[ ] $\leftarrow$ getLevelsSpanned(deviceLevels[ ], dt)
    // Get categories based on the levels spanned
    deviceCategories[ ] $\leftarrow$ getCategories(levelsSpanned[ ])

4.3.8 Example Categorization of Clients

We now use our methodology to provide an example of how we categorize the clients shown in Figure 4.1. To categorize a client we first assign a level to every RSSI value of the client using Table 4.1. We then examine the level of the signal strength (RSSI) using the median RSSI value and the levels spanned by RSSI values for different dispersion thresholds. Table 4.3 shows the results of the categorization of these clients using different dispersion thresholds. It also shows the levels of the RSSI values, of each client, using the percentiles for each dispersion threshold. For Client 1, we find that the median signal strength is in level A which corresponds to the Good category. Thus, we categorize Client 1 as Good using the median signal strength. We see from Figure 4.1 that all but one RSSI values lie in level A. Thus, for all dispersion thresholds up to the 90%, Client 1 is categorized as Good. However, using the 100% dispersion threshold, it spans level A through level C as the RSSI values for the 100th percentile are in level C. Since it spans three levels (A, B, and C) it is categorized as Moderate. Similarly, since every RSSI value for Client 2 is in level A, Client 2 is categorized as Good for all dispersion thresholds. Signals from Client 3 are received by the AP with a variety of signal strengths. Since the median signal strength of Client 3 is in level A, we categorize Client 3 as Good using the median RSSI value. Using a dispersion threshold of 50%, Client 3 spans levels A and B and the category is Good. Similarly for different dispersion thresholds up to 90% it spans two levels A and B and hence it is categorized as Good. However, using the 100% dispersion threshold the RSSI values spans three levels, A through C. Thus, using the 100% dispersion threshold it is categorized as Moderate. Similar to Client 3, the signals from Client 4 are also received by the AP with a variety of signal strengths. Using the median RSSI value and dispersion thresholds from 50% to 70% Client 4 is categorized as Good. However, since the RSSI values span 3 levels, A through C, using dispersion thresholds from 80% to 100%, Client 4 is categorized as Moderate using those dispersion thresholds. Such an analysis helps us capture the variability in signal strengths.

Dispersion Threshold	Percentile & Category	Client 1	Client 2	Client 3	Client 4
0%/ Median	50	A	A	A	B
	Category	Good	Good	Good	Good
50%	25	A	A	A	A
	75	A	A	B	B
	Category	Good	Good	Good	Good
60%	20	A	A	A	A
	80	A	A	B	B
	Category	Good	Good	Good	Good
70%	15	A	A	A	A
	85	A	A	B	B
	Category	Good	Good	Good	Good
80%	10	A	A	A	A
	90	A	A	B	C
	Category	Good	Good	Good	Moderate
90%	5	A	A	A	A
	95	A	A	B	C
	Category	Good	Good	Good	Moderate
100%	0	A	A	A	A
	100	C	A	C	C
	Category	Moderate	Good	Moderate	Moderate

Table 4.2: Results of the example categorization of all clients in a network

From these categories we see that, using dispersion thresholds of 80% and 90%, Client 1, 2, and 3 are categorized as Good. However, Client 1 and 3 are categorized as moderate if a dispersion threshold of 100% is used. This indicates that Client 1 and 3 have mostly Good signals with fewer than 10% not falling in that category. In contrast, signals from Client 2 are always Good. Client 4, on the other hand, has mostly Good signals, with fewer than 30% of the signals falling outside that category and placing it in the Moderate category for dispersion thresholds of 80% and higher.

4.3.9 Example Categorization of a Network

To categorize a network we consider the RSSI values of all messages recorded by the access point (AP). Thus, to analyze the AP shown in Figure 4.1, in a fashion similar to our analysis of clients, we first assign a level to every RSSI value. Next, we analyze the AP using the median RSSI value and the levels spanned by RSSI values for different dispersion thresholds. Table 4.3 shows the results of the categorization of the network using different dispersion thresholds. Our analysis shows that the median signal strength is in level A so using the median RSSI value the network is categorized as Good. Analyzing the network using the 50% and the 60% dispersion thresholds the network is categorized as Good since RSSI values encapsulated using those dispersion thresholds are in level A. Similarly the network is also categorized as Good using dispersion thresholds of 70%, 80%, and 90% as the RSSI values span only two levels, A and B. However, using the 100% dispersion threshold the network is categorized as Moderate, since the RSSI values spans levels A through C. These results show that for most part this AP sees signals with Good strength with fewer than 10% of the signals falling outside of that category.

Dispersion Threshold	**Percentile & Category**	**Network**
0%/ Median	**50**	A
	Category	Good
50%	**25**	A
	75	A
	Category	Good
60%	**20**	A
	80	A
	Category	Good
70%	**15**	A
	85	B
	Category	Good
80%	**10**	A
	90	B
	Category	Good
90%	**5**	A
	95	B
	Category	Good
100%	**0**	A
	100	C
	Category	Moderate

Table 4.3: Results of the example categorization of a network (AP)

Chapter 5

Analysis of Clients

In this chapter we analyze the data from all 2,946 clients in our dataset using the methodology described in Chapter 4. This is done using the RSSI values (signal strengths) of the messages received by the access points from the clients.

5.1 Categorization using the 100% Dispersion Threshold

We begin by working through an example categorization of all clients using the 100% dispersion threshold. Recall that the dispersion thresholds specify the amount of variability we encapsulate in our analysis. Comparing the results obtained using different dispersion thresholds helps us to understand the variability in the RSSIs. When analyzing this data using the 100% dispersion threshold we consider the levels assigned to the RSSI values obtained using the 0th percentile and the 100th percentile. In doing so, we encapsulate 100% of the data. Table 5.1 shows the percentage of clients in each level and category when we analyze the RSSI values of all clients in our dataset using the 100% dispersion threshold.

We see in Table 5.1 that all RSSI values for, 41.7% of the clients fall in level A, 18.4% of the clients span levels A and B, and 1.9% of the clients fall in level B. We also see in Table 5.1 that those clients in level A, A-B, and B are all categorized as having Good signal strength, and we thus have 62.0% of the clients in the Good category. Similarly, 22.2% of the clients in level A-B-C, and B-C are categorized as having Moderate signal strengths. Further, clients in level C, B-C-D, C-D, and D are categorized as having Weak signal strengths and thus 3.6% of all clients are in the Weak category. Finally, the clients whose RSSI values are spread across every level, that is A-B-C-D, are considered having Variable signal strengths, and thus 12.2% of all clients are categorized as having Variable signal strengths. In a similar fashion, we next categorize and analyze all clients using different dispersion thresholds.

Levels Spanned	Percentage of Clients in Levels	Category	Percentage of Clients in Category
A A-B B	41.7 18.4 1.9	Good	62.0
A-B-C B-C	20.0 2.2	Moderate	22.2
C B-C-D C-D D	1.4 1.0 0.7 0.5	Weak	3.6
A-B-C-D	12.2	Variable	12.2
Total	**100.0**		**100.0**

Table 5.1: Results showing the percentage of clients in different levels and categories when we analyze all clients using the 100% dispersion threshold

5.2 Analysis using Different Dispersion Thresholds

Table 5.2 shows the results of our categorizations of all clients using different dispersion thresholds. That is, it shows the percentage of clients in each of the four categories (i.e., Good, Moderate, Variable, and Weak). The difference in the percentage of clients in different categories, found using different dispersion thresholds, helps us understand the nature of the variability of the data.

Categorization of the clients							
Category	**Median**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Good	93.2%	88.5%	86.7%	85.1%	82.8%	79.0%	62.0%
Moderate	0.0%	6.1%	7.9%	9.1%	11.0%	14.2%	22.2%
Variable	0.0%	0.2%	0.3%	0.7%	1.5%	2.8%	12.2%
Weak	6.9%	5.2%	5.2%	5.1%	4.5%	3.9%	3.6%

Table 5.2: Percentage of clients in each category

For the purpose of simplifying the terminology, in subsequent discussions, clients having Good signals implies that the messages from those clients are received by the APs with Good signal strengths (RSSIs). From our analysis of clients, summarized in Table 5.2, we make three main observations.

First, the percentage of clients in the Good category is pretty high for dispersion thresholds up to 90%. Even when analyzed using the 90% dispersion threshold about 80% of the clients have Good signals and 93.2% of clients have either Good or Moderate signal strengths. This indicates that most clients have Good signals for most of the time. Comparing the results of our analysis using the 100% dispersion threshold and the 90% dispersion threshold, we find that only fewer than 10% of the messages from 38.0% of clients do not have Good signals. In other words, fewer than 10% of the signals from those clients are received by the respective APs with Moderate, Variable, or Weak signal strengths.

Second, when analyzed using the 100% dispersion threshold, we find that signals from 84.2% of the clients are received by the APs with either Good or Moderate signal strength. Thus, for the majority of the clients signal strengths are always quite reliable.

Finally, the percentage of clients categorized as Weak ranges from 6.9%, when analyzed using the median RSSI value, to 3.6% when analyzed using the 100% dispersion threshold. This means that there are a small number of clients that experience Weak signals for most of the time. In Chapter 6 we will see that these clients can have a significant influence on the way we think about and categorize the networks in which they operate.

5.3 Analysis of Clients Based on Spectrum

As explained in Chapter 4 our approach to categorization depends on both the RSSI and the receiver sensitivity. Since the receiver sensitivity also depends on the spectrum used, we next analyze the RSSI values of the messages from all clients recorded by the APs based on the spectrum used. Table 5.3 shows the categorization of all clients based on the RSSI values of the signals using the 2.4 GHz spectrum and Table 5.4 shows the categorization of all clients based on the RSSI values of the signals using the 5 GHz spectrum. We now examine the difference between signals strengths obtained when clients use the 2.4 GHz and the 5 GHz spectrum.

Category	**Median**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Good	93.3	89.9	88.2	87.2	85.7	84.2	74.8
Moderate	0.0	3.4	4.9	5.3	6.4	7.0	12.2
Variable	0.0	0.6	0.9	1.7	2.1	3.2	7.6
Weak	6.7	6.0	6.0	5.9	5.7	5.5	5.4

Table 5.3: Percentage of clients in each category using the 2.4 GHz spectrum

Category	**Median**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Good	89.2	84.6	82.6	80.9	77.7	73.2	55.3
Moderate	0.0	7.7	9.9	11.6	14.5	19.1	29.1
Variable	0.0	0.1	0.2	0.5	1.2	2.4	10.6
Weak	10.8	7.6	7.4	7.1	6.4	5.3	5.1

Table 5.4: Percentage of clients in each category using the 5 GHz spectrum

First, we observe from Table 5.3 and Table 5.4 that across all categories and dispersion thresholds, clients using the 2.4 GHz spectrum have signals about as good as or better than clients using the 5 GHz spectrum. This is as expected because of the differences in signal attenuation as described in Section 2.1.3.

Second, we observe from both Table 5.3 and Table 5.4 that there is a significant decrease in the percentage of clients categorized as Good when comparing those values when using the 90% dispersion threshold and the 100% dispersion threshold. This is similar to our earlier observation (in Table 5.2) when we analyzed the data without separating them based on the spectrums used. However, we find that the difference is more pronounced in the case of clients using the 5 GHz spectrum (the difference is nearly 20%). We observe that only 55.3% of clients using the 5 GHz spectrum compared with 74.8% of clients using the 2.4 GHz spectrum have Good signals 100% of the time. On one hand, it shows that such differences exist only for fewer than 10% of the observations. On the other hand, it shows that while the 5 GHz spectrum offers a higher rate for data transmission and is less affected by interference from sources like microwaves, and baby monitors, a significantly smaller fraction of clients operate with Good signal strengths using the 5 GHz spectrum. Irrespective of the advancements in signal encoding and WiFi protocols which promise high data rates, the messages sent by a device must be received by the receiver with Good or Moderate signal strengths in order to achieve those high rates and translate those benefits to observable improvements in the WiFi experience.

Finally, using the 100% dispersion threshold we find that messages from 5.4% of clients using the 2.4 GHz spectrum and 5.1% of clients using the 5 GHz spectrum are always received by the APs with Weak signal strengths. These clients can send messages using only the lowest two MCS indices for the messages to have a chance of being decoded by the APs. Similarly, messages received by the APs with Variable signal strengths have no certainty that they can always decode messages received using high rates. Thus, if we analyze the clients whose messages are received by the access points with Weak or Variable signal strengths, we find that 13.0% of all clients using the 2.4 GHz spectrum, and 15.7% of clients using the 5 GHz spectrum always have either Weak or Variable signals. We describe these clients as clients with unreliable signals and we analyze these devices further in Chapter 6 (when we analyze networks). The major downside of having clients on a network that can only use low PHY rates is further amplified by the WiFi anomaly problem, discussed in Section 2.1, wherein a client transmitter using low PHY rates will adversely affect other devices on all networks in the vicinity.

We note that, in the results of our analysis of clients using the 2.4 GHz spectrum (Table 5.3) and 5 GHz spectrum (Table 5.4), the percentage of clients categorized as Weak is higher when compared to the percentage of Weak clients in Table 5.2 when considered overall (without separating the signals using the 2.4 GHz and the 5 GHz spectrums). This is because the number of clients considered to calculate the percentage values in all these cases differ. While we have a total of 2,946 clients, 1,853 use only the 2.4 GHz spectrum and 1,865 use only the 5 GHz spectrum, while 782 of them use both.

Chapter 6

Analysis of Networks

To analyze a network, unlike our analysis of clients where we consider the RSSI values from only an individual client, we consider the RSSI values of messages recorded by an access point from all of its clients. In this thesis we use the terms networks and access points interchangeably, since, a network refers to an access point and every client connected to it. Similar to our analysis of clients in Chapter 5, we analyze the networks using the methodology described in Chapter 4. We first classify each RSSI value, from every client in a network, into one of four levels (A, B, C, and D) and then using the different dispersion thresholds, we categorize each of the 446 networks into one of four categories (Good, Moderate, Weak, and Variable). Table 6.1 summarizes the results of this analysis.

Category	Median	50%	60%	70%	80%	90%	100%
Good	99.3	94.8	92.0	88.4	81.1	70.2	11.3
Moderate	0.0	4.9	7.9	10.6	17.1	25.8	31.8
Variable	0.0	0.0	0.0	0.9	1.6	3.6	56.5
Weak	0.6	0.2	0.2	0.2	0.2	0.4	0.4

Table 6.1: Percentage of networks in each category using different dispersion thresholds

6.1 Analysis using Different Dispersion Thresholds

The analysis of networks using the median RSSI value (0% dispersion threshold) reveals that 99.3% of all access points (APs) receive messages with Good signal strengths from its clients and only 0.6% of APs receive messages with Weak signal strengths. A Good signal strength implies that the access point can potentially receive messages at most data rates. However, as discussed in Chapter 4 an analysis using the median value is only representative of the central tendency of the data set, and not the variability. Therefore, to capture the variability in our analysis of the networks, similar to our analysis of the clients we analyze the data using different dispersion thresholds and make three main observations.

First, we observe a general trend in Table 6.1 that as we increase the dispersion threshold in our analysis (from 50% to 100%), the percentage of the clients in the Good category decreases and the percentage of the clients in the Moderate category increases. This is because, increasing the dispersion threshold increases the amount of data encapsulated in our analysis to include more variability. Similar to the percentage of networks in the Moderate category, the percentage of networks in the Variable category also increases as we increase the dispersion threshold in our analysis.

In the previous chapter we described unreliable clients as ones whose signals are received by the AP with either Weak or Variable signal strengths. Thus our second observation from Table 6.1 is that the percentage of networks with unreliable clients increases from 4.0% using the 90% dispersion threshold to 56.9% using the 100% dispersion threshold. This demonstrates the importance of and the reason we include and examine different dispersion thresholds. While the percentage of networks in each category changes between different dispersion thresholds from 50% to 100%, the biggest changes occur between the 90% and 100% dispersion threshold. For the 90% dispersion threshold signals for 96.0% of the networks are quite reliable. However, when considering 100% of the data points, 56.9% of the networks receive unreliable signals from clients.

Finally, if we postulate that networks that receive signals with Weak or Variable signal strengths have room for improvement, we then note from Table 6.1 that using the 80% dispersion threshold only 1.8% of the networks have room for improvement. This presents a very positive but perhaps flawed narrative of the state of most networks.

Unlike a client, a network is not a singular entity. It is a collection of clients connected to an access point which must coordinate to access the shared medium. One concern with our previous analysis is that the classification of networks may be skewed by several clients that have very strong signals and only one client that has Variable or Weak signals. Thus we next analyze the networks based on the presence of unreliable clients within each network.

6.2 Networks Containing Unreliable Clients

The WiFi performance anomaly is a well-known problem affecting the WiFi networks. As discussed in Section 2.1, if a device transmits messages at slow rates, the effective throughput of other devices on the networks in the vicinity are also affected. As discussed in Chapter 4, an access point that receives messages with Weak signal strengths can only decode messages transmitted using slow rates due to the hardware limitations imposed by the receiver's sensitivity. In contrast, messages received by the AP from clients with either Good or Moderate signal strengths will be able to fairly consistently use higher rates. However, the same cannot be said about clients whose messages are received by the AP with either Variable or Weak signal strengths (unreliable clients).

Thus, we next analyze the networks with respect to the number of clients on a network whose messages are received by the access point with Weak or Variable signal strengths. These clients are forced to frequently send messages using low rates and in our analysis we describe these clients as having unreliable signals (unreliable clients). These clients are particularly important because when transmitting using low rates they can potentially reduce the effective throughput of all other clients and as a result affect the entire network.

Table 6.2 shows the percentage of the networks with one or more clients whose signals are received by the AP with unreliable signal strengths. Unlike the results of our analysis of networks in Section 6.1 (summarized in Table 6.1) which shows that the majority of the networks receive messages with Good or Moderate signal strengths, we see from Table 6.2 that across all different dispersion thresholds, more than 27.1% of all networks have one or more clients with unreliable signals. This implies that there are a considerable number of networks where the performance may be significantly hindered because one or more clients are operating with unreliable signals and as a result are forced to use slower rates. One way to interpret this finding is that there maybe a substantial number of networks that could potentially be improved. For instance, by moving clients closer to the AP, or by moving the AP closer to the clients.

Threshold	**Percentage**
Median	33.2%
50%	27.1%
60%	27.4%
70%	28.5%
80%	30.5%
90%	31.8%
100%	59.2%

Table 6.2: Percentage of networks with at least one clients with unreliable signals

We observe from our analysis of clients in Table 5.2, that using the 80% dispersion threshold, 6.0% of the clients have unreliable signals; and while analyzing the networks using the 80% dispersion threshold (Table 6.2) we find that 30.5% of networks have one or more clients with unreliable signals. In other words 179 clients with unreliable signals are distributed among 134 networks. We now further examine how these clients are distributed among networks of different sizes.

The number of networks of different sizes varies widely in our dataset. For example, while there is only 1 network of size 24, 25, and 29, there are 58 networks of size 5. This poses a challenge to our analysis as we do not want to skew the results of the analysis by using widely different numbers of networks of different sizes. To overcome this, we create groups of networks of different size ranges where each group contains roughly the same number of networks (about 25% each). The created groups contain networks that are progressively larger in size.

Figure 6.1 shows the CDF of all network sizes. Network sizes are determined by the total number of clients that connect to an access point during the entire 24 hour period. We observe from Figure 6.1 that we can group networks into four different size ranges each containing about

25% of the total networks. That is, we group networks of size 1 to 3 together which account for 25.3% of the networks. Networks of size 4 and 5 together account for 22.4%, networks of size 6 to 8 account for 28.3%, and finally networks of size greater than or equal to 9 account for 24.0% of the networks.

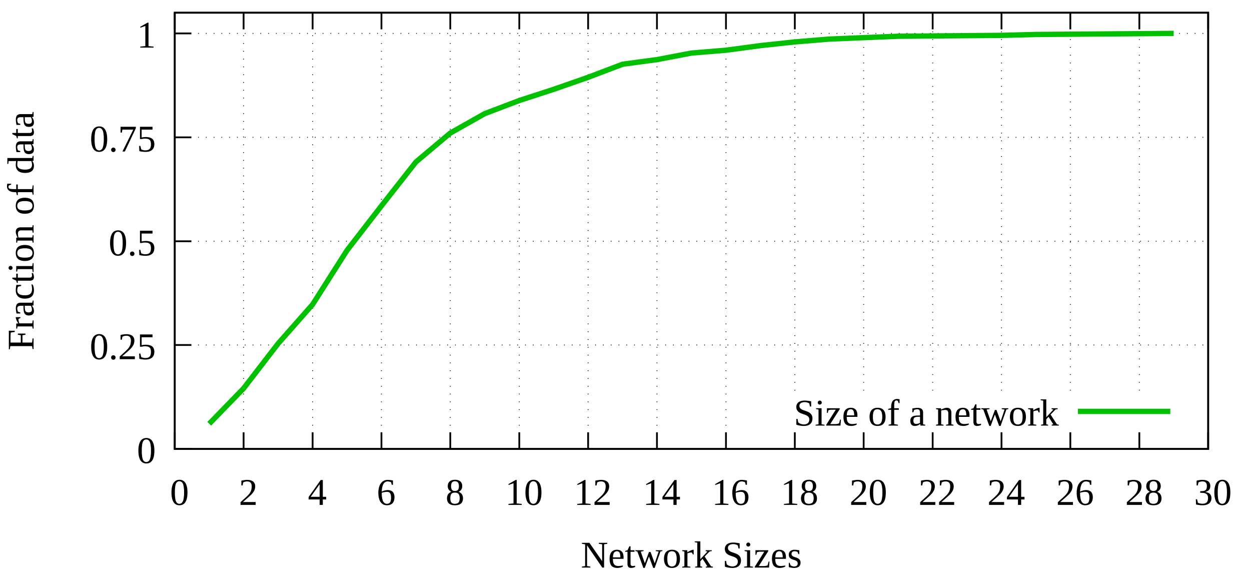


Figure 6.1: CDF of network sizes

Now that we have about the same percentage of networks in each of the different size ranges, we next analyze if the signal strengths from clients changes with the number of clients in the network. To do this, we compare the CDFs of RSSIs for access points that belong to different network size ranges. Figure 6.2 shows that the CDF of signal strengths for access points of different sizes are very similar, except for the region between -25 dBm and -45 dBm. RSSI values greater than -45 dBm are strong enough to decode messages sent at the highest rates which possibly use complex encoding formats across all frequencies (2.4 GHz and 5 GHz) and channel widths (20 MHz, 40 MHz and 80 MHz). For signal strengths less than -45 dBm, networks of different size ranges have similar trends for the CDF of RSSI values. This might lead one to believe that network sizes do not have much of an impact on the signal strengths received at the access points from the clients.

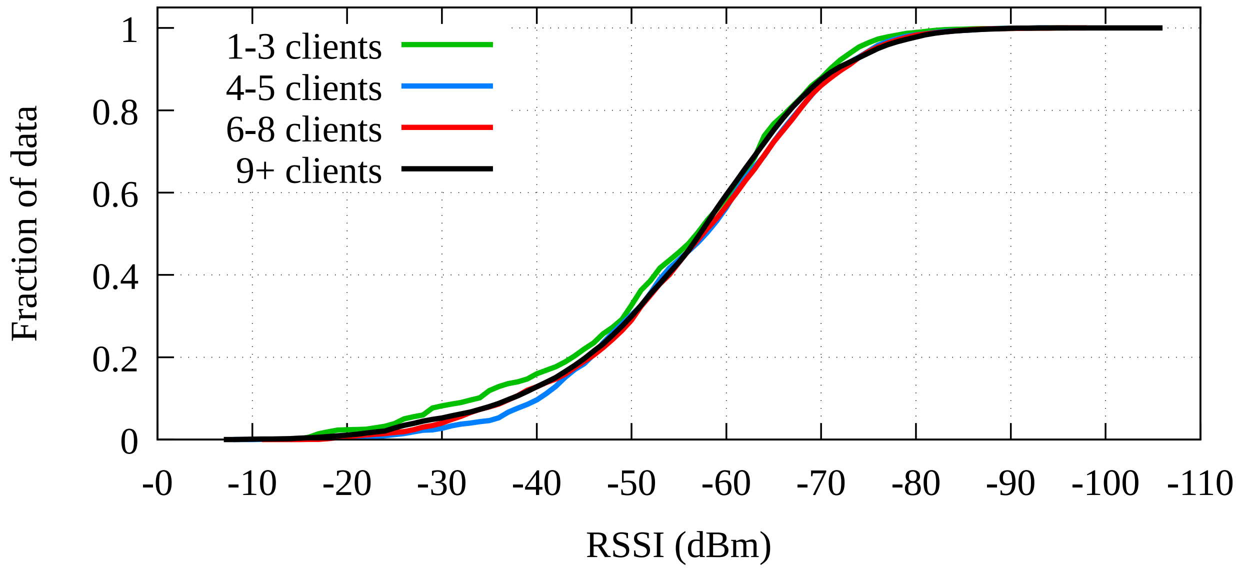
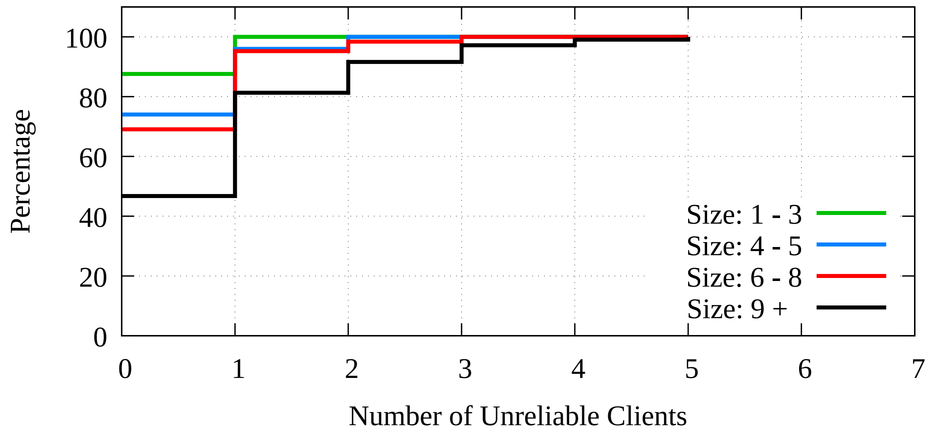


Figure 6.2: CDF of RSSI values for different network sizes

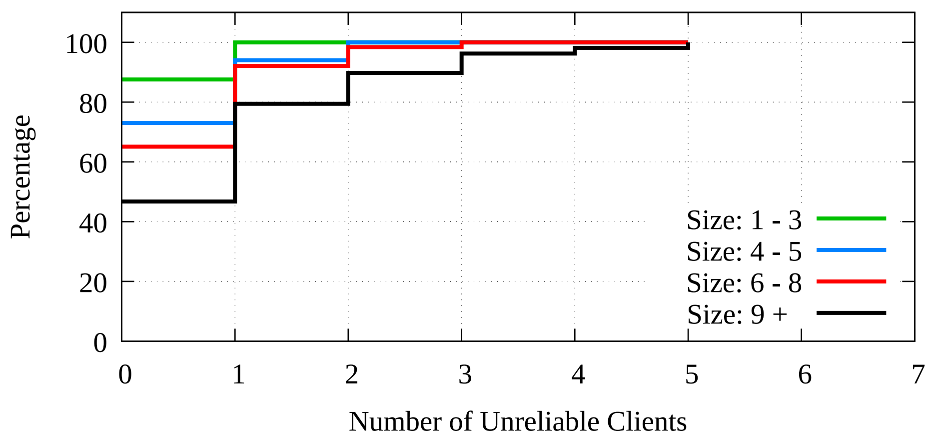
However, just based on probability one would intuitively expect networks with more clients to have more unreliable clients. Thus, an interesting question that we examine later is, are the actual fraction of networks with unreliable clients different from what the probabilities would suggest for networks with more clients? The more clients there are with unreliable signals the greater the potential for problems due to performance anomalies.

6.3 Analyzing the Impact of the Size of the Network

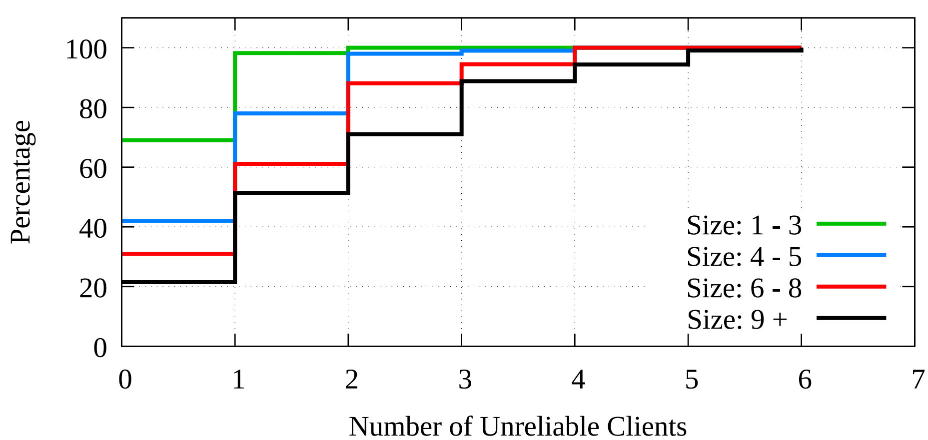
Figure 6.3a shows the CDF of the number of unreliable clients in networks of different size ranges (1-3, 4-5, 6-8, and 9 or more) using the 80% dispersion threshold. Figure 6.3b shows these results using the 90% dispersion threshold, and Figure 6.3c shows these results using the 100% dispersion threshold. When analyzing the CDFs in Figure 6.3 we make two main observations.



(a) CDF of unreliable clients using 80% dispersion threshold



(b) CDF of unreliable clients using 90% dispersion threshold



(c) CDF of unreliable clients using 100% dispersion threshold

Figure 6.3: CDFs of unreliable clients in networks of different sizes

First, we observe from all three CDFs that as the network sizes increase, the count of unreliable clients within the network increases. For instance when we analyze the CDF created using the 80% dispersion threshold (Figure 6.3a), no networks of size 1-3, have more than 1 unreliable client. However, 4.0% of the networks of size 4-5 have 2 unreliable clients, 4.8% of the networks of size 6-8 have 2 unreliable clients, and 1.6% have 3 unreliable clients. In networks of size 9 or greater, 10.3% of the networks have 2 unreliable clients, 5.6% have 3 unreliable clients, 1.9% have 4 unreliable clients, and 0.9% have 5 unreliable clients.

Second, the CDFs provide interesting insights regarding the distribution of networks with one or more unreliable clients, that are shown in Table 6.2. We note from Figure 6.3a that 12.3%, 26.0%, 30.9%, and 53.3% of the networks have one or more clients with unreliable signals in networks of sizes 1-3, 4-5, 6-8, and 9 or more, respectively. On average 30.5% of all networks have one or more clients with unreliable signals (using a dispersion threshold of 80%). Networks with more clients have a higher fraction of networks with one or more unreliable clients when compared to networks with fewer clients. However, one would intuitively expect this based on probabilities. Therefore, as mentioned previously, we next examine if the observed fraction of networks with unreliable clients is different from what the probabilities would suggest for networks of different size ranges.

To determine the expected fraction of networks with unreliable clients in different networks of different sizes (1-3, 4-5, 6-8, and 9 or more), we first determine the probability that networks of each size has 1 or more unreliable clients. Later, we use these probabilities to compute the expected fraction of networks with unreliable clients in different network size ranges. For instance, to determine the expected fraction of networks with unreliable clients, in networks of size 1-3, we first calculate the probabilities that networks of size 1, 2, and 3 have one or more unreliable clients. We then use these values to determine the expected fraction of networks with unreliable clients in networks of size 1-3.

From our previous analysis of clients in Chapter 5 summarized in Table 5.2, we can determine the percentage of clients with unreliable signals (Weak or Variable) across all networks using different dispersion thresholds. For instance, using the 80% dispersion threshold we find that, across all networks, 6.0% of the clients have unreliable signals. Thus, if one were to choose a client at random, then the probability that the particular client will have unreliable signals is 0.06. The percentage of reliable and unreliable clients found using different dispersion thresholds is shown in Table 6.3.

Dispersion Thresholds	**Median**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Percentage of unreliable clients	6.9%	5.4%	5.5%	5.8%	6.0%	6.7%	15.8%
Percentage of reliable clients	93.1%	94.6%	94.5%	94.2%	94.0%	93.3%	84.2%

Table 6.3: Percentage of clients with unreliable and reliable signals

For the purposes of our discussion, henceforth, we refer to networks with one or more unreliable clients as unreliable networks. Conversely, we refer to networks with no unreliable clients (signals from every client in the network is reliable) as reliable networks. As shown in Equation 6.1, the probability that a network of size n is unreliable ($P_u(n)$), is 1 minus the probability that the network is reliable ($P_r(n)$).

$$P_u(n) = 1 - P_r(n) \tag{6.1}$$

We assume that the probability that a single client is reliable is independent of other clients being reliable [53]. This is a reasonable assumption since clients do not impact each other's signal strengths. If the probability that an individual client in a network is reliable is p_r, then the probability that a network of size n is reliable can be determined using Equation 6.2.

$$P_r(n) = (p_r)^n \tag{6.2}$$

We use Equation 6.1 to compute the probability that a network of exactly size n has one or more unreliable clients, for networks of all sizes using different dispersion thresholds. Table 6.4 summarizes these results for dispersion thresholds of 80%, 90%, and 100%. The table shows the number of clients (n), the number of networks of that size (Count), and the fraction of the total number of networks ($f(n)$) there are of that size. In addition, it shows the expected probability of having one or more unreliable clients in networks of different sizes using different dispersion thresholds, denoted as $P_u(n)_{80}$, $P_u(n)_{90}$, and $P_u(n)_{100}$. We also include the observed fraction of unreliable networks (i.e., networks with one or more unreliable clients) using those dispersion thresholds, denoted as $O(n)_{80}$, $O(n)_{90}$, and $O(n)_{100}$.

n	Count	$f(n)$	$P_u(n)_{80}$	$O(n)_{80}$	$P_u(n)_{90}$	$O(n)_{90}$	$P_u(n)_{100}$	$O(n)_{100}$
1	27	0.061	0.06	0.04	0.07	0.04	0.16	0.19
2	38	0.085	0.12	0.16	0.13	0.16	0.29	0.34
3	48	0.108	0.17	0.15	0.19	0.15	0.40	0.35
4	42	0.094	0.22	0.21	0.24	0.19	0.50	0.48
5	58	0.130	0.27	0.29	0.29	0.33	0.58	0.66
6	48	0.108	0.31	0.23	0.34	0.29	0.64	0.67
7	47	0.105	0.35	0.36	0.38	0.43	0.70	0.74
8	31	0.070	0.39	0.35	0.43	0.32	0.75	0.65
9	21	0.047	0.43	0.52	0.46	0.52	0.79	0.81
10	14	0.031	0.46	0.29	0.50	0.43	0.82	0.71
11	12	0.027	0.49	0.33	0.53	0.33	0.85	0.58
12	13	0.029	0.52	0.46	0.56	0.54	0.87	0.69
13	14	0.031	0.55	0.29	0.59	0.36	0.89	0.93
14	5	0.011	0.58	0.60	0.62	0.40	0.91	0.60
15	7	0.016	0.60	0.86	0.65	0.71	0.92	1.00
16	3	0.007	0.63	1.00	0.67	1.00	0.94	0.67
17	5	0.011	0.65	1.00	0.69	0.80	0.95	1.00
18	4	0.009	0.67	0.75	0.71	0.50	0.95	0.75
19	3	0.007	0.69	1.00	0.73	1.00	0.96	1.00
21	3	0.007	0.73	1.00	0.77	1.00	0.97	1.00
24	1	0.002	0.77	0.00	0.81	0.00	0.98	0.00
25	1	0.002	0.79	1.00	0.82	1.00	0.99	1.00
29	1	0.002	0.83	1.00	0.87	1.00	0.99	1.00

Table 6.4: Probabilities that networks of different sizes have one or more unreliable clients

We now know both the expected probability and the observed fraction of networks with one or more unreliable clients in networks of different sizes. However, since the number of networks of each size (i.e., sample sizes) are quite small in some cases and also varies widely for different network sizes, we cannot use these values to draw reasonable conclusions. Recall that this is the reason we create groups of networks of different sizes (i.e, 1-3, 4-5, 6-8, and 9 or more). The ranges were chosen to have roughly the same number of networks in each group. Our goal is to compare the fraction of networks with one or more unreliable clients among networks of different size ranges with the expected fraction of networks with unreliable clients. To do this, we use Equation 6.3 to calculate the probability that networks in different size ranges have one or more unreliable clients. In Equation 6.3, $P_u(s, e)$ is the expected probability that networks in the size range s to e (inclusive) are unreliable.

$$P_u(s,e) = \frac{\sum_{n=s}^{e} P_u(n) f(n)}{\sum_{n=s}^{e} f(n)} \tag{6.3}$$

Table 6.5 shows both the probabilities, denoted as $P_{u_{80}}$, $P_{u_{90}}$, and $P_{u_{100}}$, and the observed fraction, denoted as O_{80}, O_{90}, and O_{100}, of networks with unreliable clients for different size ranges (1-3, 4-5, 6-8, and 9 or more) using dispersion thresholds of 80%, 90% and 100%.

Network Sizes (s to e)	Count	Percentage of Networks	$P_{u_{80}}$	O_{80}	$P_{u_{90}}$	O_{90}	$P_{u_{100}}$	O_{100}
1 to 3	113	25.30	0.13	0.12	0.14	0.12	0.31	0.31
4 to 5	100	22.40	0.25	0.26	0.27	0.27	0.54	0.58
6 to 8	126	28.30	0.35	0.31	0.38	0.35	0.69	0.69
9 or more	107	24.00	0.54	0.53	0.58	0.53	0.87	0.79

Table 6.5: Probabilities that networks of different size ranges have one or more unreliable clients

We observe from Table 6.5 that using dispersion thresholds of 80%, 90% or 100% the expected fraction and the observed fraction of networks with unreliable clients are very similar in 11 out of the 12 cases. We also note that for networks with 9 or more clients and using the 100% dispersion threshold, the fraction of observed networks is slightly lower than expected. So, overall networks with more clients are no more and no less likely to have one or more unreliable clients than expected based on probabilities. Had there been a higher fraction than expected, one reason might have been that networks with more clients might be covering a larger area (e.g., they might be deployed in larger homes or office environments). However, our findings suggest that either this is not the case or the APs that are deployed in environments that cover a larger area are part of a mesh network and clients switch to another AP when they are far enough away from the AP for which we have data.

6.4 The Case for Improving Networks

Our analysis of networks reveals that there are a considerable number of networks of all sizes that have one or more clients with unreliable signals. From Table 6.2 we see that networks with one or more unreliable clients range from 27.1% using the 50% dispersion threshold to 31.8% using the 90% dispersion threshold. We believe that all of these networks could benefit from improvements. Additionally, we observe from Table 6.3 that clients with unreliable signals range from 5.4% using the 50% dispersion threshold to 6.7% using the 90% dispersion threshold. Thus, we note that a small number of clients may impact the performance of a considerably large number of networks. The importance of our approach lies in that we examine the health of the entire network.

Finally, our analysis shows that, as expected, networks with more clients have a higher probability of having one or more unreliable clients. What is interesting is that regardless of the dispersion threshold used, the percentage of networks having one or more unreliable clients is at least 27.1% of all networks studied. Thus, in light of the WiFi anomaly problem, it seems it would be important to try to improve the signals for clients with unreliable signals. Unreliable signals may be an indication that if they are part of a mesh network some clients may not switch to another AP in the mesh network with a better signal soon enough or that some of the networks could be improved by either adding one or more APs to the mesh; or moving one or more clients or APs to improve the signals for more devices. Therefore, one avenue we hope to explore in future research is whether or not networks can be improved by having users move clients or access points to improve the overall signal quality for all clients or by adding additional APs .

Chapter 7

Conclusions and Future Work

7.1 Thesis Summary

Advances in communication technology and the proliferation of smart devices in personal, public and business spaces have led to the growth of WiFi networks. The 802.11n and 802.11ac standards enable WiFi communication between two wireless devices, a sender and a receiver. An important aspect of WiFi networks that are of concern is the maximum throughput that can be obtained. While throughput depends to a large extent on the range of PHY transmission rates supported by the client it also depends on the environments in which clients operate. That is, when a client uses a fast rate, if the strength of the received signal is not greater than the receiver sensitivity, then the receiver will not be able to decode the signal and the transmitting client will infer from packet losses that it must send messages at lower PHY rates. Failures can occur because the environment in which the devices operate can adversely affect the signal strength due to factors like reflection, absorption and attenuation from objects in the environment.

The focus of this thesis is on analyzing the signal quality of 2,946 clients operating in 446 networks on the basis of the RSSI (signal strength) of messages received by the access points to which they are connected. We characterize the signal quality to distinguish good signals from weak signals, on the basis of their strength and the range of rates that could be used during transmission. We design a methodology for categorizing clients and networks into four distinct categories Good, Moderate, Variable and Weak (using the receiver sensitivity of the chipset for a particular PHY rate). We now provide a summary of the dataset that we analyze in this thesis, the methodology we use and the key findings from our analysis of the clients and networks.

7.1.1 The Dataset

The data we analyze in this thesis was collected over a 24 hour period by sampling 446 access points, every 5 minutes, that provide connections for 2,946 clients. All data is anonymized and cannot be traced back to any particular access point, client device, or user. We examine the RSSIs of the messages from clients received by the APs since this data can be easily obtained for all clients by the AP.

The APs are equipped with the Qualcomm's IPQ 4019 system on a chip which supports simultaneous dual band communication on 2.4 GHz and 5 GHz spectrums. We also note that this chipset is used in many other commercial WiFi devices like Netgear Orbi, Linksys Velop AC2200, ASUS ZenWiFi AC (CT8) and TP-Link Deco M9. This shows the popularity of the chipset in commercial devices.

Our dataset includes a total of 417,122 data points of which 45.1% are from signals using the 2.4 GHz spectrum and the remaining 54.9% are from the 5 GHz spectrum. We find from our analysis that across all access points, the average number of clients that are simultaneously connected in any 5 minute window is quite small. That is, 65.7% of the APs have on average 3 or fewer clients that are simultaneously connected across all 5 minute windows. However, we also find that 6.5% of the APs service on average 9 or more clients.

7.1.2 Methodology for Categorization

In this thesis we design a methodology for categorizing both clients and networks, using the receiver sensitivity of the AP. We define four distinct categories (Good, Moderate, Variable and Weak). These are designed and chosen to provide an intuitive notion of better signal strength in one category over another. In addition, our methodology helps us to capture and examine the variability in signal strengths.

To categorize a client or a network we begin by first assigning a level (A, B, C, or D) to every RSSI value, for messages received by the AP from the clients. The levels indicate the range of rates that can be successfully decoded by the AP for different RSSIs using different combinations of MCSIs, spectrums and channel widths. This is an important step because the signal strength required to decode a packet depends on these factors. Signals in level 'A' are strong enough to decode all available rates. Level 'B' includes RSSI values strong enough to decode all but the rates with the highest MCS indices. Level 'C' includes RSSI values that permit the decoding of only rates using the lower 50% of the MCS indices and finally level 'D' can decode only the lowest 2 MCS indices. The levels are based on the receiver sensitivity of the wireless chipset used in the AP (Qualcomm's IPQ 4019 in our case).

Analyzing the data set, we find that over the 24 hour period, RSSIs of signals from most clients span multiple levels. Levels by themselves are insufficient to capture and analyze the variability of the RSSIs over time. Therefore, to capture this notion of variability and the range of levels that the RSSIs span we define four categories and use dispersion thresholds to categorize both the clients and the networks. The four categories we define are: Good, Moderate, Variable and Weak which have been chosen to be intuitive, with a sense of natural order among them. With signals in higher categories, it is more likely the AP will be able to decode messages sent at higher rates. To consider variability we use different dispersion thresholds and group signals into different categories for each dispersion threshold. To determine the category for a client or a network we examine the levels spanned by the RSSIs using different dispersion thresholds. As a result, we are also able to characterize the environments in which the devices operate.

We note that our methodology is general enough so it could be used with other devices with different receiver sensitivities. This simply requires changing the RSSI values used for levels A, B, C and D based on the receiver sensitivity indices for the particular AP.

7.1.3 Analysis of Clients

To analyze the clients we use our methodology to categorize them into one of four categories (Good, Moderate, Variable and Weak) using different dispersion thresholds. We find from our analysis of clients that, over the 24 hour period, 90% of the signals from 84.2% of the clients are received by the APs with either Good or Moderate signal strengths. Thus, for the majority of the clients signal strengths are mostly quite reliable. We also find that the percentage of clients categorized as Weak ranges from 6.9%, when analyzed using the median RSSI value, to 3.6% when analyzed using the 100% dispersion threshold. This means that there are a small number of clients that experience Weak signals for half to all of the time. From our analysis of the signals strengths obtained from clients that use the 2.4 GHz and the 5 GHz spectrum respectively, we find that clients using the 2.4 GHz spectrum have signals about as good as or better than clients using the 5 GHz spectrum.

7.1.4 Analysis of Networks

Similar to our analysis of clients, to analyze the networks we use our methodology and categorize the networks into one of four categories (Good, Moderate, Variable and Weak) using different dispersion thresholds. We find that when analyzing networks using dispersion thresholds between 50% and 90% only between 0.2% and 4.0% of the networks are categorized as either Weak or Variable. In other words, our analysis reveals that 90% of signals from 96.0% of networks are either Good or Moderate.

While these results sound very encouraging, a network is not a singular entity but is rather a collection of clients connected to an AP which must coordinate to access the shared medium. Therefore we analyze the networks with respect to the number of clients in a network whose messages are received by the access point with Weak or Variable signal strengths. These clients are forced to frequently send messages using low rates and in our analysis we describe these clients as unreliable because they have unreliable signals. We find from our analysis that that across all different dispersion thresholds, more than 27% of all networks have clients with unreliable signals. This implies that there are a considerable number of networks where the performance may be significantly hindered because one or more clients are operating with unreliable signals and as a result are forced to use slower rates. We also find that networks with more clients have more clients with unreliable signals and that the fraction of networks with one or more clients with unreliable signals is quite close to what is expected based on probabilities.

From our analysis of networks we see that networks with unreliable clients range from 27.1% using the 50% dispersion threshold to 31.8% using the 90% dispersion threshold. We believe that many of these networks could benefit from improvements. Additionally, we observe from our analysis of clients, that clients with unreliable signals range from 5.4% using the 50% dispersion threshold to 6.7% using the 90% dispersion threshold. Thus, we note that a small number of clients may impact the performance of a considerably large number of networks. What is interesting is that regardless of the dispersion threshold used, the percentage of networks with unreliable clients is at least 27.1% of all networks studied. The importance of our approach lies in that we examine the health of the entire network.

7.2 Future Work

We now briefly describe some possible avenues for future work.

7.2.1 Obtaining and Analyzing a Larger Dataset

Our dataset was obtained from 2,946 clients operating in 446 networks over a period of 24 hours sampled every 5 minutes. It would be interesting to obtain data for a longer period of time (e.g., weeks or months) and for more networks and clients. It would also be interesting to determine if the results are sensitive to the 5 minute sampling period by obtaining data from a more frequent sampling period. Interesting future research would also examine similarities and differences across a variety of regions (countries and subcontinents). For example, while the data we analyze in this thesis was obtained on a weekday, by analyzing a week's worth of data we can compare weekday and weekend usage. Likewise, analyzing a month's worth of data might provide insights into network usage during different seasons and holidays.

7.2.2 Analyzing Mesh Networks

Our study highlights that there is significant room for improvement in many WiFi networks. Previous studies have shown that mesh networks can help improve WiFi networks [6]. Our results show that networks with more clients tend to have more clients from which the signals received by the AP are unreliable. One interesting avenue of future work would be to obtain data from mesh deployments to see if they have improved signals from clients.

7.2.3 Improving Existing Networks

In this thesis, we design a methodology to analyze clients and networks using their signal strengths. It would be interesting to study if it is possible to suggest that users move one or more of their devices that operate using Variable or Weak signal strengths closer to the AP or, alternatively move the AP in order to improve the overall signal quality for more devices. Another possible recommendation that may help improve signal quality would be to suggest the addition of another node to form a mesh network.

7.3 Concluding Remarks

The proliferation of wireless devices in both business and consumer spaces coupled with the rise of IoT enabled devices has increased our reliance on WiFi networks. Demand for more WiFi networks with high performance will only increase. It is therefore important that the messages sent by clients are received with strong enough signal strength to decode the messages. Unfortunately, we find that as network sizes increase, the number of clients with unreliable signals also increases and these clients will be restricted to lower rates which can negatively impact all other clients on the network.

Previous studies have characterized WiFi networks using the measured throughput of the clients. However, the throughput experienced and the rates used by clients in those studies depend on the capability of the clients. If the client does not support the 5 GHz spectrum or is limited to one spatial stream its range of possible PHY rates will be limited. The significant advantage of our approach is that it is independent of the capabilities of the clients used in the study. We instead analyze the potential range of rates than can be used based solely on the RSSI values of messages arriving at the AP. Our study provides insight into possible benefits from deploying clients with newer technologies. Only networks where most of its clients are operating with Good signals will be able to see significant improvements with newer technologies, as clients with Weak signals will not be able to use fast rates under any technology.

Finally, we draw attention to our methodology for categorization where we use the ability of the chipset to decode messages sent using a particular rate which depends on the signal strength of the received message. Our methodology can be adapted to characterize any WiFi networks using any chipset. In addition, the categories chosen by us are intuitive which roughly translates to the range of physical rates that can be used by the devices.

References

[1] Alyaa Syaza Azini, Muhammad Ramlee Kamarudin, and Muzammil Jusoh. Transparent antenna for WiFi application: RSSI and throughput performances at ISM 2.4 GHz. *Telecommunication Systems*, 61(3):569–577, 2016.

[2] Anand Balachandran, Geoffrey M. Voelker, Paramvir Bahl, and P. Venkat Rangan. Characterizing User Behavior and Network Performance in a Public Wireless LAN. In *ACM SIGMETRICS Performance Evaluation Review*, volume 30, pages 195–205. ACM, 2002.

[3] Dinesh Bharadia and Sachin Katti. Full Duplex MIMO Radios. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI '14)*, pages 359–372, 2014.

[4] Helmut Bolcskei, David Gesbert, and Arogyaswami J. Paulraj. On the Capacity of OFDM-Based Spatial Multiplexing Systems. *IEEE Transactions on Communications*, 50(2):225–234, 2002.

[5] Ranveer Chandra, Ratul Mahajan, Thomas Moscibroda, Ramya Raghavendra, and Paramvir Bahl. A Case for Adapting Channel Width in Wireless Networks. In *ACM SIGCOMM Computer Communication Review*, volume 38, pages 135–146. ACM, 2008.

[6] Aizaz U. Chaudhry, Roshdy H. M. Hafez, Osama Aboul-Magd, and Samy A. Mahmoud. Throughput Improvement in Multi-Radio Multi-Channel 802.11a-Based Wireless Mesh Networks. In *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, pages 1–5. IEEE, 2010.

[7] Cisco. Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper, Feb 2020.

[8] Gary Davis and Ralph Jones. The Sound Reinforcement Handbook (Yamaha), 1990.

[9] Lara Deek, Eduard Garcia-Villegas, Elizabeth Belding, Sung-Ju Lee, and Kevin Almeroth. The Impact of Channel Bonding on 802.11n Network Management. In *Proceedings of the Seventh Conference on Emerging Networking Experiments and Technologies*, page 11. ACM, 2011.

[10] Lara Deek, Eduard Garcia-Villegas, Elizabeth Belding, Sung-Ju Lee, and Kevin Almeroth. Intelligent Channel Bonding in 802.11n WLANs. *IEEE Transactions on Mobile Computing*, 13(6):1242–1255, 2013.

[11] Ning Ding, Daniel Wagner, Xiaomeng Chen, Abhinav Pathak, Y. Charlie Hu, and Andrew Rice. Characterizing and Modeling the Impact of Wireless Signal Strength on Smartphone Battery Drain. *ACM SIGMETRICS Performance Evaluation Review*, 41(1):29–40, 2013.

[12] David Eckhardt and Peter Steenkiste. Measurement and Analysis of the Error Characteristics of an In-Building Wireless Network. In *ACM SIGCOMM Computer Communication Review*, volume 26, pages 243–254. ACM, 1996.

[13] Zhen Fang, Zhan Zhao, Daoqu Geng, Yundong Xuan, Lidong Du, and Xunxue Cui. RSSI Variability Characterization and Calibration Method in Wireless Sensor Network. In *The 2010 IEEE International Conference on Information and Automation*, pages 1532–1537. IEEE, 2010.

[14] Sandra Fiehe, Janne Riihijärvi, and Petri Mähönen. Experimental Study on Performance of IEEE 802.11n and Impact of Interferers on the 2.4 GHz ISM Band. In *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, pages 47–51, 2010.

[15] Market Research Firm. Wi-Fi Market by Component and Services, Density and Region - Global Forecast to 2022, Aug 2019.

[16] Matthew Gast. *802.11 Wireless Networks: The Definitive Guide*. O'Reilly Media, Inc., 2005.

[17] Dirk Gates. 5 WiFi Networking Predictions For 2017. https://www.networkcomputing.com/wireless-infrastructure/5-wifi-networking-predictions-2017, Jan 2019.

[18] Leo A. Goodman. Kolmogorov-Smirnov Tests For Psychological Research. *Psychological bulletin*, 51(2):160, 1954.

[19] E. S. Gopi. Modulation Techniques in Wireless Communication. In *Digital Signal Processing for Wireless Communication using Matlab*, pages 103–169. Springer, 2016.

[20] IEEE 802.11 Working Group et al. IEEE Standard For Information Technology–Telecommunications and Information Exchange Between Systems–Local And Metropolitan Area Networks–Specific Requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments. *IEEE Std*, 802(11), 2010.

[21] Sarthak Grover, Mi Seon Park, Srikanth Sundaresan, Sam Burnett, Hyojoon Kim, Bharath Ravi, and Nick Feamster. Peeking Behind the NAT: An Empirical Study of Home Networks. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, pages 377–390. ACM, 2013.

[22] Mohamed Hadi Habaebi and Nur Izzati Nabilah Bt Azizan. Harvesting WiFi Received Signal Strength Indicator (RSSI) For Control/Automation System In SOHO Indoor Environment with ESP8266. In *2016 International Conference on Computer and Communication Engineering (ICCCE)*, pages 416–421. IEEE, 2016.

[23] Habanero. IPQ 4019 Datasheet. https://www.8devices.com/media/products/habanero/downloads/Habanero_datasheet.pdf. Accessed: 23-1-2020.

[24] Lajos Hanzo, Yosef Akhtman, Jos Akhtman, Li Wang, and Ming Jiang. *MIMO-OFDM for LTE, WIFI and WIMAX: Coherent versus Non-Coherent and Cooperative Turbo-Transceivers*. John Wiley & Sons, 2011.

[25] Martin Heusse, Franck Rousseau, Gilles Berger-Sabbatel, and Andrzej Duda. Performance Anomaly of 802.11b. In *IEEE INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, volume 2, pages 836–843. IEEE, 2003.

[26] Kenichi Higuchi, Akihiro Fujiwara, and Mamoru Sawahashi. Multipath Interference Canceller for High-Speed Packet Transmission With Adaptive Modulation and Coding Scheme in W-CDMA Forward Link. *IEEE Journal on Selected Areas in Communications*, 20(2):419–432, 2002.

[27] W. Cary Huffman and Vera Pless. *Fundamentals of Error-Correcting Codes*. Cambridge University Press, 2010.

[28] Lito Kriara, Mahesh K. Marina, and Arsham Farshad. Characterization of 802.lln Wireless LAN Performance via Testbed Measurements and Statistical Analysis. In *2013 IEEE International Conference on Sensing, Communications and Networking (SECON)*, pages 158–166. IEEE, 2013.

[29] Tianbo Kuang and Carey Williamson. RealMedia Streaming Performance on an IEEE 802.11b Wireless LAN. In *Proceedings of IASTED Wireless and Optical Communications (WOC)*, pages 306–311, 2002.

[30] Chi-Min Li and Hsueh-Jyh Li. An Dual-Polarized Transmission System with OFDM Multiplexing. In *APMC 2001. 2001 Asia-Pacific Microwave Conference (Cat. No. 01TH8577)*, volume 2, pages 609–612. IEEE, 2001.

[31] Zixin Liu, Fuhai Li, Yihong Qi, and Ji Chen. An Effective Receiver Sensitivity Measurement. In *2015 IEEE Symposium on Electromagnetic Compatibility and Signal Integrity*, pages 310–313. IEEE, 2015.

[32] Gough Lui, Thomas Gallagher, Binghao Li, Andrew G. Dempster, and Chris Rizos. Differences in RSSI Readings Made by Different WiFi Chipsets: A Limitation of WLAN Localization. In *2011 International Conference on Localization and GNSS (ICL-GNSS)*, pages 53–57. IEEE, 2011.

[33] Cisco Networks. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper, Feb 2019.

[34] Konstantina Papagiannaki, Mark Yarvis, and W. Steven Conner. Experimental Characterization of Home Wireless Networks and Design Implications. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, pages 1–13. Citeseer, 2006.

[35] Ashish Patro, Srinivas Govindan, and Suman Banerjee. Observing Home Wireless Experience through WiFi APs. In *Proceedings of the 19th Annual International Conference on Mobile Computing and Networking*, pages 339–350. ACM, 2013.

[36] Konstantinos Pelechrinis, Theodoros Salonidis, Henrik Lundgren, and Nitin Vaidya. Experimental Characterization of 802.11n Link Quality at High Rates. In *Proceedings of the Fifth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization*, pages 39–46, 2010.

[37] Eldad Perahia and Robert Stacey. *Next Generation Wireless LANs: 802.11n and 802.11ac*. Cambridge University Press, 2013.

[38] Chuan-Chin Pu and Wan-Young Chung. Indoor RSSI Characterization using Statistical Methods in Wireless Sensor Network. *The Journal of The Korea Institute of Maritime Information & Communication Sciences*, 11(11):2172–2178, 2007.

[39] Rafhanah Shazwani Rosli, Mohamed Hadi Habaebi, and Md. Rafiqul Islam. Characteristic Analysis of Received Signal Strength Indicator from ESP8266 WiFi Transceiver Module. In *2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, pages 504–507. IEEE, 2018.

[40] Tony J. Rouphael. *RF and Digital Signal Processing for Software-Defined Radio: A Multi-Standard Multi-Mode Approach*. Newnes, 2009.

[41] Nurul I. Sarkar and Kevin W. Sowerby. The Combined Effect of Signal Strength and Traffic Type on WLAN Performance. In *2009 IEEE Wireless Communications and Networking Conference*, pages 1–6. IEEE, 2009.

[42] Warren L. Stutzman and Gary A. Thiele. Gain Calculations for Reflector Antennas. *Antenna Theory and Design*, pages 433–436, 1981.

[43] Warren L. Stutzman and Gary A. Thiele. *Antenna Theory and Design*. John Wiley & Sons, 2012.

[44] Srikanth Sundaresan, Sam Burnett, Nick Feamster, and Walter De Donato. BISmark: A Testbed for Deploying Measurements and Applications in Broadband Access Networks. In *2014 USENIX Annual Technical Conference (USENIXATC 2014)*, pages 383–394, 2014.

[45] Srikanth Sundaresan, Nick Feamster, and Renata Teixeira. Measuring the Performance of User Traffic in Home Wireless Networks. In *International Conference on Passive and Active Network Measurement*, pages 305–317. Springer, 2015.

[46] Markus Tauber and Saleem N. Bhatti. Low RSSI in WLANs: Impact on Application-Level Performance. In *2013 International Conference on Computing, Networking and Communications (ICNC)*, pages 123–129. IEEE, 2013.

[47] Ajay Tirumala, Les Cottrell, and Tom Dunigan. Measuring end-to-end bandwidth with Iperf using Web100. In *In Web100, Proceedings of Passive and Active Measurement Workshop*. Citeseer, 2003.

[48] Mike J. Usher and D. A. Keating. *Sensors and Transducers: Characteristics, Applications, Instrumentation, Interfacing*. Macmillan International Higher Education, 1996.

[49] Lochan Verma, Mohammad Fakharzadeh, and Sunghyun Choi. WiFi on Steroids: 802.11ac and 802.11ad. *IEEE Wireless Communications*, 20(6):30–35, 2013.

[50] Jean Walrand and Pravin Pratap Varaiya. *High-Performance Communication Networks*. Morgan Kaufmann, 2000.

[51] Peter J. Winzer. Making Spatial Multiplexing a Reality. *Nature Photonics*, 8(5):345, 2014.

[52] Baoguo Yang, Khaled Ben Letaief, Roger S. Cheng, and Zhigang Cao. Timing Recovery for OFDM Transmission. *IEEE Journal on Selected Areas in Communications*, 18(11):2278–2291, 2000.

[53] Lotfi Asker Zadeh. Probability Measures of Fuzzy Events. *Journal of Mathematical Analysis and Applications*, 23(2):421–427, 1968.

[54] Yunze Zeng, Parth H. Pathak, and Prasant Mohapatra. A First Look at 802.11ac in Action: Energy Efficiency and Interference Characterization. In *2014 IFIP Networking Conference*, pages 1–9. IEEE, 2014.

[55] Weihao Zhou, Zhi Wang, and Wenwu Zhu. Mining Urban WiFi QoS Factors: A Data Driven Approach. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 9–16. IEEE, 2017.