U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

A Framework for Usage Modeling and Anomaly Detection in Large-Scale 802.11 Networks

Dossa Mohamed Massa

Tese de Doutoramento Apresentada

à Faculdade de Engenharia da Universidade do Porto em

Telecommunications Engineering

Supervisor: Prof. Ricardo Santos Morla (PhD)

April 2015

Abstract

Wireless 802.11 networks are a popular technology that offers inexpensive ubiquitous access to the Internet in campuses, enterprises, homes, coffee shops, airports, and other public places. Their wide-scale adoption has brought great convenience to many people, giving them anytime and anywhere access to the Internet. As a result people are becoming more and more dependent on these networks and are increasingly demanding reliability and high performance when connecting to the Internet through them. Due to the inherent problems of the wireless medium, these demands from the users pose significant challenges for network administrators. Learning how each part of the network behaves and how it is used is fundamental in addressing these challenges. As a network grows in scale and usage, these demands become an even greater challenge and it may not be feasible for administrators to use the same techniques that are used for small deployments. It is increasingly difficult for network administrators to maintain knowledge of infrastructure usage properties including usage patterns of individual access points (APs), users, and locations as well as their susceptibility to different problems as a deployment scales up. Network administrators require other complementary techniques for network management.

In this thesis, we propose to use the analysis of collected 802.11 network usage data to aid performance and fault management of large-scale networks. Realistic knowledge of the usage patterns of 802.11 networks is critical for performance management, in a bid to make sure that resources are provisioned according to usage and that the network delivers the desired performance with respect to the expected usage. Learning from the collected 802.11 data is also crucial for fault management, as it may help network administrators to efficiently detect and fix different connectivity and performance problems facing the users of large-scale 802.11 networks. To simplify management of large-scale 802.11 networks, in this thesis we contribute with a framework for 1) usage modeling and 2) anomaly detection, both based on the analysis of collected 802.11 usage data.

Most previous works on usage modeling focus on the user perspective of 802.11 including user mobility, registration, dwelling, and encounter patterns. In our framework we propose and evaluate a number of probabilistic models for automatic characterization of access point (AP) usage. We include time-dependent and time-independent models of AP usage characterization, as well as models that consider AP week structure usage namely weekdays, weekends, and individual days of the week (Monday-Sunday).

On the other hand, most previous works on anomaly detection uses enhanced devices (clients and APs), hardware sensors, sniffers, and controllers for detecting anomalies such as interference, overload, and halted or crashed APs. In our framework, we propose a methodology for detecting patterns of AP usage anomaly based on the analysis of the relationship between session endings at 802.11 AP. We identify a usage pattern named “abrupt ending” of 802.11 AP connections that happens when a large number of user sessions in the same access point (AP) end within a one second window. We propose an algorithm for automatic detection and characterization of different anomaly-related patterns associated to AP abrupt ending occurrences. We confirm the existence of significant statistical relationship between abrupt ending occurrences, anomaly-related patterns occurrences, and aggregate 802.11 network usage in terms of total number of sessions. We confirm the existence of abrupt endings in other 802.11 deployments. Our findings indicate abrupt endings are primarily the effect of interference across the 802.11 infrastructure and usage pattern behavior of the APs, but also misconfiguration and bugs on the 802.11 APs. Users and their respective device’s specifics play no significant role in abrupt ending manifestation. We finally provide an online implementation of the detection and characterization of abrupt endings and their respective anomaly-related patterns using the Esper complex event driven processing engine.

“The more you know about your 802.11 network, the more you can do with it and the better you can make it”

John Cox: Best practices for managing WLANs (2008)

Resumo

As redes sem fio 802.11 são uma tecnologia popular que oferece acesso ubíquo e barato à Internet no campus, na empresa, em casa, no café, no aeroporto, e noutros espaços públicos. A sua ampla adoção é conveniente para muitas pessoas, dando-lhes acesso à Internet em qualquer altura e em qualquer lugar. Como resultado desta adoção, as pessoas estão a tornar-se cada vez mais dependentes destas redes e exigem cada vez mais fiabilidade e alto desempenho quando se ligam à Internet através destas redes. Devido aos problemas inerentes do meio sem fios, estas exigências dos utilizadores colocam desafios significativos para os administradores destas redes. Aprender como cada parte da rede se comporta e como é utilizada é fundamental para endereçar estes desafios. À medida que a rede cresce em escala e utilização, estas exigências dos utilizadores tornam-se um ainda maior desafio e pode não ser viável os administradores utilizarem as mesmas técnicas que utilizam para pequenas redes. É cada vez mais difícil para os administradores de rede manterem conhecimento das propriedades da utilização da infraestrutura - incluindo padrões de utilização de pontos de acesso sem fio (APs) individuais, utilizadores, e localizações bem como a sua susceptibilidade a diferentes problemas à medida que a rede se torna maior. Os administradores de rede precisam de outras técnicas complementares para a gestão da rede.

Nesta tese propomos utilizar a análise de dados de utilização recolhidos da rede 802.11 para ajudar à gestão de desempenho e falhas em redes de grande escala. O conhecimento realista dos padrões de utilização das redes 802.11 é crítico para a gestão de desempenho, de modo a garantir que os recursos são aprovisionados de acordo com a utilização e que a rede oferece a capacidade desejada para a utilização esperada. Aprender a partir dos dados recolhidos da rede 802.11 é também crucial para a gestão de falhas, já que pode ajudar os gestores de rede a detetar e resolver eficientemente problemas de conectividade e desempenho que afetam os utilizadores de redes 802.11 em grande escala. De modo a simplificar a gestão destas redes, nesta tese contribuímos com uma framework para 1) modelização da utilização e 2) detecção de anomalias, ambas baseadas na análise de dados de utilização recolhidos da rede 802.11.

A maioria dos trabalhos anteriores em modelização da utilização dá ênfase à perspetiva do utilizador incluindo padrões de mobilidade, registo, estadia, e encontro. Na nossa framework propomos e avaliamos vários modelos probabilísticos para a caracterização automática da utilização de APs. Incluímos modelos dependentes e independentes do tempo, bem como modelos que consideram a estrutura semanal da utilização dos APs nomeadamente dias da semana, fins de semana, e dias individuais da semana (Segunda-feira a Domingo).

Por outro lado, a maioria dos trabalhos anteriores em detecção de anomalias utiliza dispositivos aumentados (clientes e APs), sensores em hardware, sniffers, e controladores para detetar anomalias como interferência, sobrecarga, e APs parados ou em falha. Na nossa framework propomos uma metodologia para detetar padrões de anomalias de utilização de APs baseada na análise de relações entre término de sessões nos APs. Identificamos um padrão de utilização chamado “término abrupto” de ligações a APs 802.11 que ocorre quando um grande número de sessões de utilizador no mesmo AP terminam na mesma janela de um segundo. Propomos um algoritmo para a detecção e caracterização automática de vários padrões anómalos associados a ocorrências de términos abruptos. Confirmamos a existência de uma relação estatística significativa entre ocorrência de términos abruptos, ocorrência de padrões relacionados com anomalias, e utilização agregada da rede 802.11 em termos de número total de sessões. Confirmamos a existência de términos abruptos em outras redes 802.11. O que descobrimos indica que os términos abruptos são principalmente o efeito de interferência através da infraestrutura 802.11 e padrões de comportamento dos APs, mas também configurações erradas e bugs nos APs. Os utilizadores e as características dos seus dispositivos não tomam um papel significativo na manifestação de términos abruptos. Por fim providenciamos uma implementação online da detecção e caracterização de términos abruptos e dos seus respetivos padrões de anomalia utilizando o motor Esper de processamento de eventos complexos.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my academic advisor Prof. Ricardo Santos Morla, whom his constant support, stimulating discussions, valuable comments and insights provided me the perfect guidance that I needed through my Ph.D. journey. Ricardo’s comments and ideas always come at the appropriate moment and place with the right amount, keeping me busy and focused always. I am very glad that all our efforts were useful and valuable. A part of the effort has turned into completion of this thesis, and the rest prepared me well for my future careers.

I would also like to thank my employer the Institute of Finance Management in Tanzania (IFM) and the Foundation for Science and Technology (FCT) in Portugal for their extended financial support throughout my Ph.D. studies. Certainly, without their support results of this work would not be the same. FCT support was received through grant number SFRH/BD/69824/2010. I am also thankful that in the context of FCT, particularly project SUM through PTDC/EIA/113999/2009, I received UMinho data set which I used for validation of my results in this thesis. Also, I am very grateful to INESC-TEC direction for allowing me to conduct my research in their highly reputable research laboratory (CTM). It's an incredibly positive working environment. In addition, I am very thankful to all my colleagues including those at IFM, in the MAP-tele program, and at INESC-TEC for their constant encouragement, support, and friendship. I am also very grateful to all my friends in Porto for the good moment we spent together. They really helped me in making this beautiful city a home away from home.

Moreover, I would like to thank all members of my family, especially my parents Hajj Mohamed .K. Massa and the late Nibaro Athuman Liku, for being a source of joy in my life and for making me a person I am today both personally and professionally. Also, I am enormously grateful to all my seven brothers, two sisters, and all my in-laws for their endless love, encouragement, advice, and support. They always encouraged me through the rough times and advised me not to give up on this Ph.D. work. I am thankful for all their efforts and I feel very lucky to have them in my family.

Last but not least, I would like to thank my wife Halima Sharifu. She has been very considerate and supportive throughout my Ph.D. process: always turning around the boring situations into laughter and depressing ones into consolations. I am specifically grateful for her encouragement to let me pursue what I dream of rather than what would have seemed convenient for us. I am also enormously grateful to our children Nibaro, Abdillah, and Abdulrahman. Nibaro came just to visit me here with my father and decided to stay until the end of my Ph.D. Little Abdillah was born at the last stage of my Ph.D. work, while a tiny Abdulrahman was born during thesis write-up phase. Indeed, they both brought much fun, happiness, and joy into our lives which are impossible to enumerate all here.

POPH QUALIFICAR É CRESCER.
QREN QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL PORTUGAL 2007.2013
GOVERNO DA REPÚBLICA PORTUGUESA
UNIÃO EUROPEIA Fundo Social Europeu

Table of Contents

List of Acronyms

AIC	Akaike Information Criterion
AP	Access Point
ARIMA	Auto-Regressive Integrated Moving Average
BSS	Basic Service Set
CDF	Cumulative Distribution Function
CORBA	Common Object Request Broker Architecture
CPD	Conditional Probability Distribution
CRC	Cyclic Redundancy Check
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
CTS	Clear to Send
DBSCAN	Density based Clustering Algorithm
DCF	Distribute Coordination Function
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name Server
DS	Distribution System
DSSS	Direct Sequence Spread Spectrum
ESS	Extended Service Set
FCAPS	Fault, Configuration, Accounting, Performance, and Security
FEUP	Faculty of Engineering of the University of Porto
FHSS	Frequency Hopping Spread Spectrum
GPS	Global Positioning System
HTTP	Hypertext Transfer Protocol
IAB	Internet Activities Board
IBSS	Independent Basic Service Set
IETF	Internet Engineering Task Force
IP	Internet Protocol

ISM	Industrial, Scientific and Medical
ISO	International Standard Organization
LL	Log-likelihood
LLC	Logical Link Control
MAC	Media Access Control
MIB	Management Information Base
MIMO	Multiple-Input and Multiple-Output
NIC	Network Interface Card
OFDM	Orthogonal Frequency Division Multiplexing
PCF	Point Coordination Function
PDF	Probability Density Function
PHY	Physical Layer
QoS	Quality of Service
RADIUS	Remote Authentication Dial-In User Service
RF	Radio Frequency
RMON	Remote Network Monitoring
RSSI	Received Signal Strength Indicator
RTS	Request to Send
RTT	Round Trip Time
SINR	Signal to Interference Noise Ratio
SMI	Structure of Management Information
SNMP	Simple Network Management Protocol
SSID	Service Set Identifier
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
U-NII	Unlicensed National Information Infrastructure
VoIP	Voice Over IP
WLAN	Wireless Local Area Network

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Management Challenges of Deployed 802.11 Networks

Wireless 802.11 networks that span across several buildings in universities and enterprises or that connect several homes and cities are referred to as large-scale networks. Large-scale networks are typically characterized by many network elements such as 802.11 access points mainly for providing coverage and capacity, for supporting a large number of users typically with heterogeneous client devices and a variety of applications, and also for enabling large volumes of traffic that arise due to the intensive use of the network at various distinct locations in the coverage area.

As 802.11 networks grow in scale and usage, the challenges for network management also grow. For a small network, it is straightforward for the administrator to acquire by experience and manual inspection the baseline knowledge of the overall infrastructure usage characteristics. Administrators may know about: 1) usage behavior of individual users, e.g. the APs they tend to associate, their average access durations, application preferences, and the amount of traffic they tend to generate; 2) individual access points usage properties, e.g. when are they busy and idle, the amount of traffic they normally process, and their susceptibility to different problems; and 3) location preferences in terms of network usage (e.g. library, cafeteria, student center, etc.) and locations that are prone to problems such as interference or overload.

In contrast, for a large-scale network this knowledge is difficult to acquire due to the physically distributed nature of the network and its high number of APs, users, and network locations. For example, consider a campus network with several hundred APs and thousands of users, spread across a large physical space, and covering about 30 buildings both indoor and outdoor. In this case, maintaining a mental baseline knowledge of the infrastructure usage characteristics including usage patterns of individual APs, users, and locations is not practical and may not be feasible. Scale poses difficulties in following up usage of APs, users, and locations in the infrastructure, let alone detecting any potential problems.

In addition, challenges associated to shared spectrum and the time varying nature of the wireless medium occur both in small and large 802.11 networks. Wireless channel conditions continuously evolve over time and change differently in different parts of the environment. This means that at any given moment in time one or more access points and users in certain locations of the infrastructure are likely to be facing problems such as interference, intermittent connections, or authentication failure. These problems can affect multiple APs and locations at the same time in the infrastructure and can be difficult to detect given the scale of the network. Having a device walk/run near every AP for detection of problems in the network proves to be not only inconvenient but also expensive.

To constantly ensure that users of the network have consistent access, that problems on the network can be quickly identified and resolved, and that the network provides desired capacity and reliably, 802.11 administrators require complementary approaches such as learning from the collected 802.11 network usage data.

1.2 Trace-Based Analysis for 802.11 Network Management

Network management encompasses a wide range of activities as depicted by the ISO FCAPS model [15, 16], which defines five areas of network management functions: fault, configuration, accounting, performance, and security. Analysis of the collected network usage data can play a crucial role towards fulfilling some of the goals of these management functions, given the aforementioned complexities of large-scale 802.11 networks. The following discusses how this can be achieved in each of the five management areas. 1) Fault management: different abnormal network usage behaviors of APs, users, and locations can be detected by observing different indicators of abnormality in the data e.g. increased packet error rate or packet retransmissions rate. 2) Configuration management: problems related to device misconfigurations (APs or users) in any location of the network can be detected by observing appropriate patterns in the usage data e.g. authentication failure for legitimate users or frequently disconnected clients. 3) Accounting management: tracking network usage e.g. association durations and amount of bytes/packet sent and received of individual users, group of users, departments or units is possible irrespective of their locations. 4) Performance management: Over-utilized/under-utilized APs and locations in terms of usage in the infrastructure can be observed and together with their corresponding times of the day e.g., their aggregate association durations, number of sessions established, and traffic. 5) Security management: unauthorized access to a network as well as unauthorized access points attached to a network in any location of the network can be identified from the trace data, e.g. unknown/unregistered user and AP MAC addresses.

While configuration, accounting, and security remain important functions of 802.11 network management; our emphasis in this thesis is on the application of trace-based analysis on performance and fault management of a large-scale network. Performance management is focused on ensuring that network performance is maintained at acceptable thresholds. To realize this goal effectively, appropriate plan and models for network usage need to be developed from trace analysis, so that resources in the network can be provisioned with proper balance to achieve the desired network performance, in relation to usage. The goal of fault management is to detect and correct faults that occur in the network. To detect and correct faults quickly and efficiently, appropriate indicators for network usage anomalies must be devised and monitored from the trace data, so that algorithms and models can be developed for identification of their true nature and for their proper characterization. Our approach of using collected

802.11 usage data analysis for performance and fault management aims to fulfill the aforementioned objectives in such a manner. In this way, administrator in charge of managing 802.11 networks can stay ahead of impending capacity and connectivity problems likely to occur in large-scale networks.

1.3 Limitation of Current Analyses

The focus of early 802.11 trace-based works was to understand general characteristics of 802.11 network usage. Numerous basic statistics about 802.11 network usage were then presented: like the average number of users, average session's duration, bytes sent and received, and protocols used [51 - 55]. Although these statistics helped to understand the underlying usage properties of 802.11 networks, there were no attempts to model this usage.

The major focus of recent studies on 802.11 networks is on user mobility [56 - 80], user encounter patterns at APs [81 - 88], user registration patterns at APs [89 - 95], user dwelling time at APs [96 - 100], and traffic [101 - 109]. These works provide broad classification of users based on their degree of mobility, encounter patterns, arrival and departures patterns to APs, and traffic. However, their focus is more on the users' perspective of network usage rather than the access point, e.g. mobility, association, and traffic models of individual users or group of users. In addition, it is not possible to derive synthetic samples from these models for using e.g. in a network simulator as most of the proposed models are not probabilistic in nature.

On the other hand, most works for anomaly detection proposed in the literature aim to detect overload and flash crowd at an AP [117 - 128], failed and disconnected clients at AP [129 - 132], anomalous signal strength variations at AP and interference [136 - 147], and rogue AP [145 - 153]. However, in order to detect anomalies most of these works suggest the use of enhanced devices (clients and APs), sniffers, and third party hardware device such as controllers/sensors rather than looking at AP usage patterns. While detecting patterns of anomaly by looking at AP usage data can be simple and effective, deploying and maintaining these devices in 802.11 infrastructures is difficult and expensive.

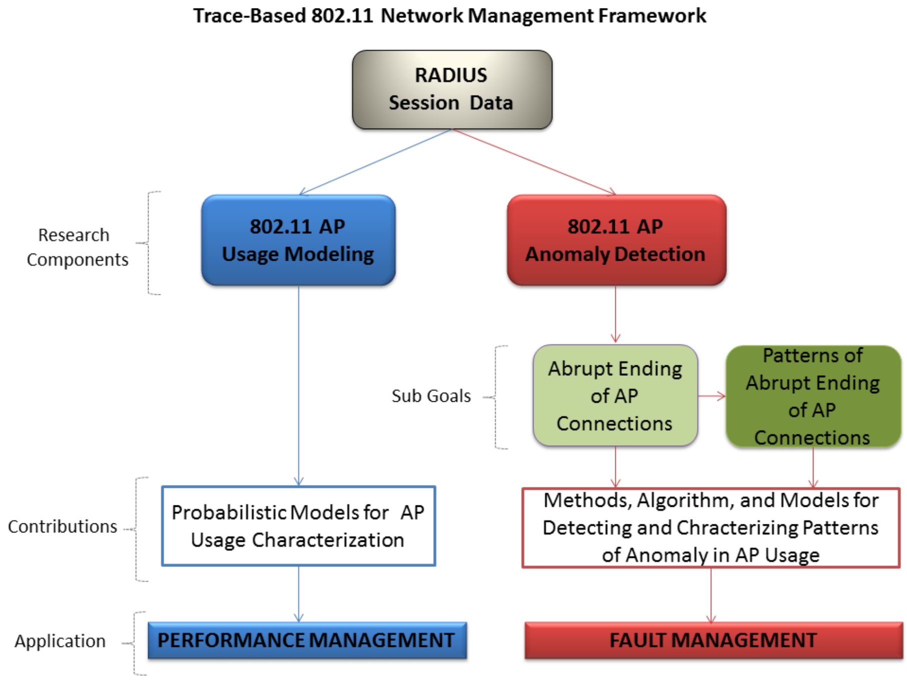
Along the same line of research there are other works that aim to detect hidden terminals, stations experiencing capture, anomalous traffic, and intrusions at AP [154 - 175]. However, the performance of each proposed 802.11 deployment scenarios in these frameworks is evaluated either based on a small scale network, on a testbed, or on simulations, thus hardly quantifying the underlying characteristics of a large-scale 802.11 deployment. There exist also several open source and commercial-based products that aid 802.11 network management tasks [176 - 179]. However, their main focus is to provide aggregate statistics and graphs of 802.11 network usage rather than detecting patterns of anomaly in AP usage and establishing their respective nature.

Current approaches do not support automatic detection and characterization of anomaly-related patterns in the usage of APs namely AP interference, AP crash, AP overload, user intermittent connectivity and authentication failure patterns at AP which makes them not viable option for immediate adoption as network management tools.

1.4 Framework for Usage Modeling and Anomaly Detection

Our focus in this thesis is on probabilistic models of AP usage and detecting patterns of anomaly in the usage of AP. The emphasis on AP usage and the use of probabilistic generative models is better suited to 1) performance management of large-scale 802.11 infrastructures, e.g. models for capacity planning suitable for resource provisioning of large-scale 802.11 networks, and 2) network simulation tools for evaluating applications and protocols that may run in 802.11 infrastructures. The emphasis on detecting patterns of anomaly in the usage of AP and establishment of their respective nature is better suited to 3) fault management of large-scale 802.11 networks, as it may help network administrators to quickly and efficiently detect and fix problems linked to usage of the large-scale 802.11 networks.

With the aforementioned 1-3 items in mind, in this thesis we propose a framework for realistic usage modeling and anomaly detection for large-scale 802.11 networks. The high level architecture of this framework is shown in figure 1.



Trace-Based 802.11 Network Management Framework
RADIUS Session Data
Research Components
802.11 AP Usage Modeling
802.11 AP Anomaly Detection
Sub Goals
Abrupt Ending of AP Connections
Patterns of Abrupt Ending of AP Connections
Contributions
Probabilistic Models for AP Usage Characterization
Methods, Algorithm, and Models for Detecting and Chracterizing Patterns of Anomaly in AP Usage
Application
PERFORMANCE MANAGEMENT
FAULT MANAGEMENT

Figure 1. Proposed framework for usage modeling and anomaly detection

Our proposed framework can be used in two network management functions as far as FCAPS is concerned, namely performance and fault management. Our proposed framework can be part of a proactive network management strategy for handling network capacity and reliability problems before they affect network and network services. For example, the performance management feature of our framework aims to provide network administrators with the ability to understand and to predict usage of APs in 802.11 infrastructures. Understanding AP usage can be helpful in resource allocations of a large-scale 802.11 network. Proper allocation of network resources may enable 802.11 networks to deliver performance according to the expected usage and traffic demand of 802.11 users, thereby resolving many of performance problems such as AP overload, degraded throughput, and delay. On the other hand, the fault management feature of our framework aims to help 802.11 administrators in diagnosing and detecting patterns of anomaly associated to usage of APs in the infrastructure. In this way, appropriate remedial actions/decisions can be taken depending on the nature of anomaly diagnosed, as an attempt to offer and guarantee consistent connectivity to wireless users. Being able to properly plan for usage of APs and to quickly and efficiently detect patterns of anomaly in their usages, allows administrators to proactively manage 802.11 networks against performance and connectivity degradations. This is to emphasize the usefulness of our proposed framework to 802.11 administrators in aiding performance and fault management tasks.

To develop our proposed framework, we use a data set collected from the underlying 802.11 environment, namely 802.11 session data collected centrally from the server running RADIUS (Remote Authentication Dial-In User Service) protocol [36]. This is a set of 802.11 traces from the Eduroam hotspot of the Faculty of Engineering of the University of Porto (FEUP). In our 802.11 network access points are configured to monitor associations of individual users in the network, and all association events are recorded in the server running RADIUS authentication protocol. In this way usage activity of individual access points and users extended over seconds, minutes, hours, days, weeks, months, and years is maintained. The collected RADIUS data can therefore be used for different purposes such as realistic 802.11 AP usage modeling and anomaly detection as proposed in this thesis.

In our framework, we investigate different methods of automatic learning from the data and methods for parameter and threshold estimation from the data. We include into our framework: 1) models for characterizing AP usage, 2) methodologies and an algorithm for detecting patterns of anomaly in AP usage, as well as models for characterizing their occurrences in relation to 802.11 network usage - both based on the collected 802.11 AP session data.

For usage modeling: we attempt to derive generative probabilistic models of AP usage based on daily keep-alive event counts. A keep-alive event is a message sent by a mobile client every 15 minutes for refreshing the client’s association with an AP. Due to their periodic nature, keep-alive statistics give us estimates of the time users stay associated with an AP during which they may generate traffic with different profiles. We decided to model keep-alive event counts given the evidence that user-AP association pattern and event counts are correlated with access duration and traffic at APs [104, 108, 113].

We therefore propose a number of generative statistical models for access point (AP) usage characterization based on daily counts of keep-alive event events and compare them using log-likelihood and Akaike Information Criterion (AIC) values figures of merit. These models include time-independent and time-dependent models for characterizing AP usage i.e., models that consider independency between consecutive usage samples of APs and those that consider dependency in time. The time-independent models investigated are a simple exponential model, a discrete mixture of exponentials, a continuous Gamma mixture of exponentials, and an above-below model that allows thinking about the usage problem in binary terms (i.e. AP usage as either high or low). The time-dependent models include a table conditional probability distribution model that uses binary variables for predicting the next value of an AP usage given its previous usage samples. We further consider week structure usage samples of AP and evaluate AP usage models: for weekdays, weekends, and individual days of the week (Monday through Sunday). We understand that for any modeling-based efforts to be effective, it is important to assess the performance of the proposed models. With this in mind, we provide cross-validation comparison results of all our models based on log-likelihood and AIC values on specific training and test data sets.

For anomaly detection: we propose a methodology for detecting patterns of AP usage anomaly and their respective nature from the analysis of AP usage data. We focus on a usage pattern named “abrupt ending” of 802.11 AP connections. Abrupt ending of connections refers to a situation where an 802.11 AP drops all or a significant part of its users’ connections within a one second time window, as seen from RADIUS authentication logs [36]. During abrupt endings, mobile stations typically change association to other APs, and it can take few seconds, minutes, or even hours before an AP starts accepting new connections again. Each time a handoff happens, management frames are exchanged between the station and the target AP, thus keeping the medium busy and preventing other stations from accessing the wireless medium [119, 123, 155]. The need to associate and re-associate, especially during abrupt ending can have significant impact on the users-AP connectivity and traffic performance.

Recognizing potential connectivity disturbance of wireless user connections upon encountering abrupt ending of 802.11 AP connections, in our framework we propose a method for detecting the abrupt ending of AP connections from collected 802.11 AP usage data. We also propose an algorithm for detecting and characterizing abrupt endings into several patterns of AP usage anomaly. We detect AP-related as well as user-related anomalous patterns. The detected AP-related patterns include: 1) interference across AP vicinity, i.e. when neighboring APs encounter abrupt endings in succession within one minute, 2) AP persistent interference, i.e. when a particular AP has encountered abrupt endings repeatedly in one day, 3) AP overload, i.e. when continued user sessions exist after an abrupt ending, meaning the abrupt ending occurred as a result of AP over-utilization by these users, 4) AP halt/crash, i.e. when no single user session/connection existed after abrupt ending during a specified time period, indicating an AP that is not running, and 5) AP interference, i.e. when abrupt ending does not belong to any of the aforementioned patterns. The detected user-related patterns are: 1) user authentication failure, i.e. inability of users to connect to an AP, which is observed when an abrupt ending resulted from a single user; and 2) user intermittent connectivity to AP, i.e. when frequently disconnected user sessions are observed at AP, particularly before and after abrupt ending. We also include in our framework statistical models for characterizing abrupt ending of AP connections occurrence and anomaly-related patterns occurrence in relation to aggregate 802.11 network usage, in terms of total number of sessions. These models include linear regression models and continuous probability models such as Exponential, Gamma, and Gaussian distributions. The proposed models can help 802.11 administrators to estimate occurrences of abrupt endings and their resulting anomaly-related patterns in the usage of APs of these large-scale 802.11 infrastructures.

1.5 Contributions

The scope of this thesis encompasses two major research thrusts: 1) proposing and evaluating models of 802.11 AP usage applicable to large-scale networks, and 2) investigating patterns of AP usage anomaly and techniques needed to detect and characterize them efficiently. Both 1) and 2) are based on the collected 802.11 AP usage data. The key contributions we make are:

- A set of probabilistic models for access point (AP) usage characterization. These include time-independent and time-dependent models for automatic 802.11 AP usage characterizations.
- Considering week structure usage samples of APs and proposing AP usage models for weekdays, weekends, and specific individual days (Monday through Sunday).
- The identification of a new usage pattern named “abrupt ending” of 802.11 AP connections from the analysis of AP usage data. We propose an algorithm for the detection and characterization of different anomaly-related patterns associated to AP abrupt ending, and we also propose statistical models for characterizing their occurrences with respect to aggregate network usage.
- The confirmation of the existence of significant statistical relationship between abrupt ending occurrences and aggregate 802.11 network usage, in terms of total number of sessions. We show this relationship is also significant for most anomaly-related patterns we investigate and can be modeled by linear regressions as well as exponential distributions.
- The confirmation that abrupt endings and their respective underlying anomaly-related patterns are the general phenomena among the 802.11 deployments we analyzed, and that they are the consequence of interference across the infrastructure, usage pattern behavior of APs, and bugs/misconfigurations of APs, while users and specifics of user devices plays no significant role in their manifestation.
- An online implementation of the detection and characterization of abrupt endings and their associated anomaly-related patterns using the Esper complex event driven processing engine.

These contributions can also be found in the following list of our publications:

- Massa, D., & Morla, R. (2013). Modeling 802.11 AP Usage through Daily Keep-Alive Event Counts. *Wireless networks, 19*(5), 1005-1022, [47].
- Massa, D., & Morla, R. (2010). Modeling 802.11 AP Usage through Daily Keep-Alive Event Counts. In *16th International Conference on Network-Based Information Systems* (pp. 195-200). IEEE, [48].

- Massa, D., & Morla, R. (2013). Abrupt Ending of 802.11 AP Connections. In *Computers and Communications (ISCC), Symposium on* (pp. 000348-000353). IEEE, [49].

In addition, we have one more paper titled “Detecting and Modeling Patterns of 802.11 AP Abrupt Ending of Connections” submitted to a journal for possible publication at a time of this thesis submission.

1.6 Thesis Structure

In chapter 2 we provide background materials on 802.11 networks. We explain their physical and protocol architectures, their inherent problems and challenges for network management. In chapter 3 we provide survey of the related work pertaining to the problem of usage modeling and anomaly detection in large-scale 802.11 networks. In chapter 4 we present our modeling work of 802.11 AP usage. We evaluate and compare different probabilistic models (time-dependent and time-independent) for characterizing AP usage based on AP daily keep-alive event counts. In chapter 5 we present the first part of our anomaly detection work. We identify a usage pattern named abrupt ending of AP connection from the analysis of 802.11 AP usage data, and we investigate their occurrences and propose models for their characterization. In chapter 6 we present an algorithm for characterizing abrupt endings as one of a set of different forms of anomaly-related patterns in the usage of 802.11 APs, and propose models for characterizing their occurrences with respect to aggregate network usage. We also include an online implementation of the detection and characterization of abrupt endings and their resulting anomaly-related patterns in this chapter. In chapter 7 we provide conclusions and suggest directions for future work.

Chapter 2

Overview of 802.11 Networks

2.1 Network Architecture

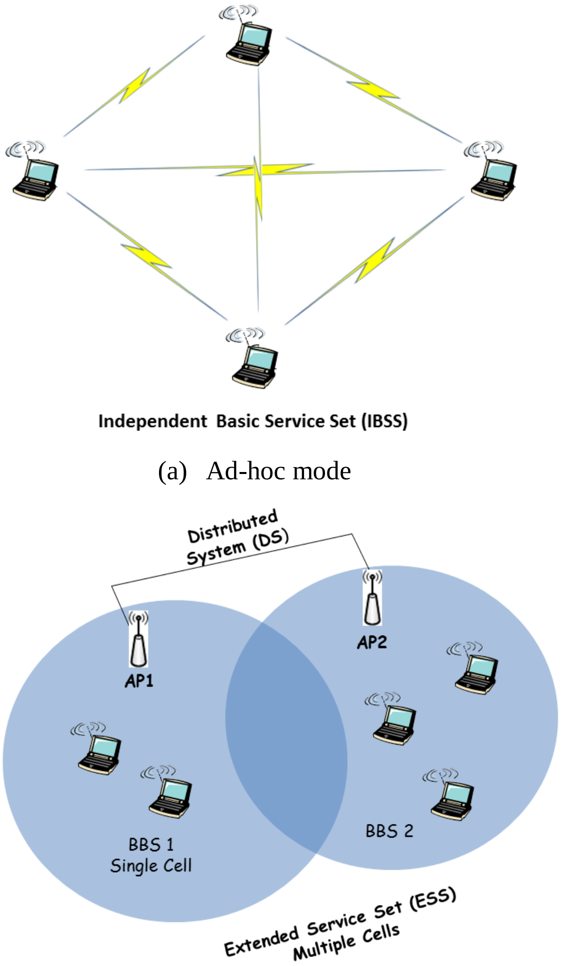
A wireless 802.11 network (WLAN or Wi-Fi) is a type of computer network that is designed to offer location-independent network access among various computing devices by means of radio waves rather than a cable infrastructure [1]. In campuses 802.11 networks are typically deployed as the final link connecting the existing backbone wired network to a group of client stations, allowing these users to roam while accessing the full resources and services of the network across buildings or campus. The most obvious motivation and benefit behind these 802.11 deployments is increased nomad behavior. This means wireless 802.11 users are no longer tied to a location. A wireless user can move about freely with their devices from one place to another and access 802.11 networks without the restrictions of connecting to the network backbone. The other benefit includes cost-effective network setup, particularly for locations which are difficult to wire, e.g. older buildings and solid-wall structures. To offer 802.11 wireless services in these areas requires only installation of base stations and antennas, rather than running cables and patching in new Ethernet jacks, while adding users is just a matter of authorization. This can also translate to reduced cost of ownership, particularly in environments which are dynamic in nature that may perhaps require frequent modifications [1].

In 802.11 networks, each device with 802.11 capabilities (whether mobile, portable, or fixed), is referred to as a wireless **station**. In fact, these wireless stations are of two categories: access points and clients [1]. Wireless **Access Points** (APs) usually act as base stations to 802.11 networks. They connect wireless clients to the existing wired network and handle any communications between them. **Wireless clients** are mobile devices equipped with a wireless adapter. These devices include PCs, laptops, PDAs, tablets, and smart phones. Usually, the wireless adapter in the client device communicates with the access point using RF signals. When connection is established, wireless clients can have access to the network and network services just as if they are part of the wired network.

When two or more stations are in communication range and happen to communicate to each other they form a **Basic Service Set (BSS)** [1]. The smallest possible BSS consists of two stations. A station in the same basic service area can communicate with other members of the BSS. A BSS that is not connected to a base station is called an **Independent Basic Service Set (IBSS)** or an **Ad-Hoc** network (figure 2(a)). In an ad-hoc network, stations communicate directly with each other in a peer to peer fashion. There is no base station and thus no device to coordinate communications. This also implies stations in ad-hoc network cannot connect to any other basic service set. Ad-hoc networks usually involve a few number of stations set up for certain objectives and for a short duration such as in a disaster recovery [2].

When two or more BSS's are interconnected they form an infrastructure network [2]. Infrastructure network uses base stations for all communications, including communication between stations in the same BSS. For communication to be achieved, all stations are required to be within the communication range of the access point. Stations must first associate to the access point in order to use the infrastructure network.

Two or more BSS's can be interconnected using a **Distribution System (DS)** [1, 2]**.** DS increases network coverage by linking access points to form an extended, larger network. This means each BSS becomes a component of an extended network, which makes seamless mobility between BSS's easier to achieve. Admission to the DS is via the use of access points, and also data moves between the BSS and the DS with the aid of these access points.



(a) Ad-hoc mode

(b) Infrastructure Mode

Figure 2. Wireless 802.11 network architecture[1]

Creating arbitrarily large and complex networks using BSS's and DS's leads to the next level of hierarchy called the **Extended Service Set (ESS)** (figure 2(b)). The interesting thing of the ESS is that the entire 802.11 network looks like an independent basic service set to the Logical Link Control layer (LLC) [2]. This means irrespective of basic service area, stations within the ESS can communicate with each other even when moving between BSS′s.

Distribution system supports the following mobility-related transitions. For example, **No-transition**: if a station is stationary or is moving only within its own BSS. **BSS-transition**: when a station moves between BSS's in the same ESS. **ESS transition**: when a station moves between BSS's belonging to different ESS's [3]. However, for a station to use 802.11 networks, it must first associate itself with the BSS infrastructure, typically via an access point. Association events are very dynamic in nature, because stations are always on the move, and are frequently turned on and off. A wireless station can only be associated with one AP at a time [3]. This allows DS to know the identity and the location of access point a station is associated with.

In addition to association, there are two other association-related services supported by DS. When a station moves between BSSs it will switch access point. This service is termed as **reassociation** [3]. Reassociation is normally initiated by the station, in particular when signal strength indicates different association can be helpful. When reassociation is complete, the DS updates its location records to reflect station reachability via a new access point. **Disassociation** service is when the existing association between the mobile station and the AP is terminated. This can be triggered by either party. Once disassociation is complete, a disassociated station can no longer send or receive data.

2.2 Protocol Architecture

2.2.1 Data Link Layer

IEEE 802.11 specifications focus on the two lowest layers of the OSI model, incorporating both physical and data link components. Each 802.11 network contains both a MAC and a Physical (PHY) component. The architecture allows multiple physical layers to be developed to support the 802.11 MAC. The MAC layer regulates how to access the medium and send data, while transmission and reception is dealt by the physical layer (PHY).

The data link layer in these 802.11 networks comprises of two sub-layers: Logical Link Control (LLC) and Media Access Control (MAC). 802.11 networks make use of the same 802.2 LLC and 48-bit addressing similar to other 802 LANs, where MAC address is unique. This allows for simple bridging between 802.11 networks and existing IEEE backbone networks. The 802.11 MAC is designed to support multiple competing nodes to share the radio medium [4].

To control access to the transmission media and for collision detection, 802.3 Ethernet LANs employ Carrier Sense Multiple Access with Collision Detection **(CSMA/CD)** protocol [4]. In this scheme, a station must first listen to the media before attempting to transmit, and once a collision is detected, a transmitting station stops its transmission of the frame, transmits a jam signal, and then waits for a random time interval before trying to resend the frame. To detect a collision, a station must have the ability to both transmit and listen at the same time, which is not the case in 802.11 network environments [5].

Rather than collision detection employed by 802.3 Ethernet (CSMA/CD), 802.11 uses a Carrier Sense Multiple Access with Collision Avoidance **(CSMA/CA)**. The 802.11 standard refers to this scheme as the Distributed Coordination Function (DCF) [5]. CSMA/CA incorporates the use of explicit acknowledgements (ACK) for all transmitted packets. This means an ACK packet is sent by the receiving station to confirm if the data packet has arrived intact. If ACK is not received by the sending station, collision is assumed to have occurred and the data packet is retransmitted after waiting for a random period of time.

Although this explicit ACK mechanism employed by 802.11 systems controls interference and other radio related problem, introduces overhead to 802.11 networks. In this manner, an 802.11 LAN will typically have slower performance in comparison to its Ethernet LAN counterpart.

The hidden node problem is another MAC-layer problem significant to wireless 802.11 networks, when some stations are out of range of other stations or a group of stations (see figure 3). In this case, two stations situated on opposite sides of an access point can both hear activity from a given access point, but not directly from each other, mostly due to a distance or an obstruction [5].

To prevent collision resulting from hidden nodes, 802.11 specifies an optional Request to Send/Clear to Send **(RTS/CTS)** protocol at the MAC layer [5]. A sending station initiates the process by sending an RTS and waits for CTS reply from the access point. Since the access point is heard by all stations in the vicinity, the CTS cause them to delay any impending transmissions, hence allowing the sending station to transmit and receive a packet acknowledgment without any chance of interference from any hidden nodes. Although RTS/CTS handle the hidden node problem, nevertheless RTS/CTS procedure adds significant overhead to the network by temporarily reserving the medium before any data transmission could commence.

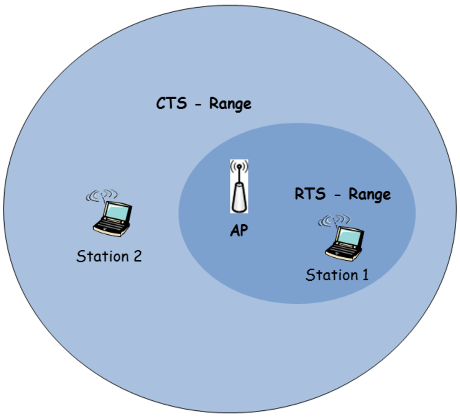


Figure 3. RTS/CTS Procedure [1]

In addition, the 802.11 MAC layer offers two more robustness features named **CRC checksum** and **packet fragmentation** [5]. Each packet has a CRC checksum calculated and attached to it. This ensures that the received data was intact. On the other hand, packet fragmentation allows for large packets to be broken into smaller units when sent over the wireless channel. This helps to improve reliability when congestion or interference is a factor, because in these situations larger packets stand better chance of being corrupted/dropped than smaller fragments [6]. Fragmentation reduces retransmissions significantly by reducing the amount of data that could be corrupted, which in turn improves overall wireless network performance. The MAC layer in the receiving side is responsible for reassembling fragments, allowing the process to be transparent to higher level protocols.

Moreover, 802.11 MAC specifications provide support for time-bounded data such as voice and video through the **Point Coordination Function (PCF)** [6]. In PCF mode a single access point controls access to the media, contrary to **Distribute Coordination Function (DCF)** where control is distributed to all stations. In PCF mode, access point polls one station after another for data in a predetermined amount of time. However, a major limitation to PCF is scalability. For an access point to have control of media access and also to poll all associated stations, this can be ineffective for a large-scale network.

2.2.2 Physical Layer

The second major component of 802.11 protocol architecture is the physical layer (PHY). The physical layer essentially provides wireless transmission mechanisms for the MAC, in addition to supporting secondary functions such as evaluating the state of the wireless medium and reporting it to the MAC [7]. The independence between the MAC and PHY has enabled the addition of higher data rates, for example 802.11b (11 Mbit/s), 802.11a (54 Mbit/s), 802.11g (54 Mbit/s) and 802.11n (600 Mbit/s).

The IEEE 802.11 standard originally describes three physical layers: an infrared layer, a frequency-hopping spread-spectrum (FHSS) layer, and a direct-sequence spread-spectrum (DSSS) layer [7]. The original 802.11 wireless standard defines data rates of 1 Mbps and 2 Mbps based on radio waves using frequency hopping spread spectrum (FHSS) or direct sequence spread spectrum (DSSS). FHSS and DSSS are basically different signaling mechanisms that cannot interoperate with each other. With frequency hopping technique, the entire 2.4 GHz band is divided into 75 1-MHz sub-channels. A hopping pattern is agreed between a sender and a receiver, and data is sent over a sequence of the sub-channels [7]. With FHSS in use, each communication takes place in a different hopping pattern within 802.11 systems. The hopping patterns are designed to minimize simultaneous use of the same sub-channel by two senders.

FHSS techniques allow for a fairly simple radio design (only precise timing is required for controlling frequency hops), however, they are limited to speeds of up to 2 Mbps. This limitation is imposed by FCC (Federal Communications Commission, USA) that confines sub-channel bandwidth to 1 MHz. These regulations drive FHSS systems to spread their usage through the entire 2.4 GHz band; meaning they must hop often which results into a high amount of hopping overhead.

The direct sequence signaling technique divides the 2.4 GHz band into 14 22-MHz channels. Adjacent channels may overlap partially. Data can be sent in any one of these 22 MHz channels without needing to hop to other channels.

A technique namely chipping is usually employed to compensate for noise on a specific channel [8]. In this case, each bit of user data is transformed into a series of redundant bit patterns known as chips. The inherent redundancy nature of each chip together with spreading the signal across the available 22 MHz channel offers a form of error checking and correction. Even if it happens that a part of the frequency band is interfered, in most cases data can still be recovered, thereby reducing the need for retransmissions [8].

The most important factor hampering adoption of 802.11 networks is limited throughput. The data-transmission rates specified by the original 802.11 standards are too slow to support most general business requirements. In order to provide support for higher data rates, the IEEE has been rectifying standards from time to time. For example, the 802.11-1997 was the first wireless standard in the family. In terms of acceptance, 802.11b was the first to be accepted broadly and then followed by 802.11a and 802.11g, while presently 802.11n is a latest multi-streaming modulation technique [9].

802.11b and 802.11g make of use of the 2.4 GHz ISM; hence their equipment’s are more susceptible to interference from devices using the same frequency bands such as microwave ovens, cordless telephones, and Bluetooth devices. To control interference, 802.11b and 802.11g equipment use direct-sequence spread spectrum (DSSS) and orthogonal frequency-division multiplexing (OFDM) signaling methods, respectively [10]. 802.11a uses 5 GHz U-NII band that provides at least 23 non-overlapping channels. This is different to the 2.4 GHz ISM frequency band, where adjacent channels are allowed to overlap [10]. Depending on the operating environment, improved or poor performance can be obtained using either higher or lower frequencies (channels). The latest 802.11n has the capability to be enabled in both the 5 GHz mode and 2.4 GHz mode, if there is knowledge about the likelihood of interference from other 802.11 or non-802.11 radio devices operating in the same frequencies. By coupling the MIMO architecture with wider-bandwidth channels allowed 802.11n to have increased transmission speed over 802.11a (5 GHz) and 802.11g (2.4 GHz) [10].

2.3 Challenges of Managing 802.11 Networks

The unique nature of the wireless medium make the management of 802.11 networks very different and more challenging than in the wired network. With wireless systems, the radio waves propagate in space and have to pass through any obstacle, matter, or concrete material that exists in the coverage area e.g. walls, floors, and ceilings. Matter can reflect, scatter, and partially absorb radio waves resulting into weak signal strength and reduced coverage area; the effect can be so severe particularly if the coverage area has lots of metals [11]. Presence of matter in the coverage area can cause dead spots (i.e. areas in which the 802.11 network simply does not work), and RF effects such as hidden terminals.

Another inherent characteristic of the wireless channel is the variation of the channel strength over time and over frequency [4]. As a consequence of the time-varying channel, wireless transmission is often prone to errors. Usually errors on a wireless link occur in long bursts, especially when the node appears to be fading out (i.e. when the received signal strength drops below a certain threshold). In addition to interference caused by broadcast nature of wireless links, where transmission in one link of the 802.11 network interferes with the transmissions in adjacent links, 802.11 wireless transmissions is affected by other radio waves operating in the same frequency range [12, 138]. The interference simply means data rarely makes it through, otherwise requiring lots of retransmissions resulting in overall poor performance.

Central to the problems of 802.11 networks is the association strategy employed by the existing 802.11 architectures, i.e. the method on how to select the best AP to associate to among the available APs. In the existing architecture, the received signal strength is the only criterion used to select an AP for association [13, 126]. This leads to scenarios where some APs in the network have fewer users, while other APs are overloaded with too many users, because their signal strengths as measured by the mobile stations appear to be strong. This can negatively impact stations throughput and overall effective use of the 802.11 network and network services.

Lastly, are the challenges revolving around the unique characteristics of users of 802.11 networks, i.e. evolving nature of their movement and associations to APs. Usually, in these large-scale 802.11 deployments the environment is not static, for example in a campus people and objects constantly move around. In addition to creating obstacles for reflection of 802.11 signals, this behavior is likely to change coverage patterns of APs and their planned capacity [14, 133]. The situation can be particularly problematic when an 802.11 infrastructure becomes densely populated. This means usage requirements for any given AP or physical location are likely to change from time to time. In these circumstances, managing a wireless 802.11 network for high reliability and performance can be a difficult task without efficient and robust models. In fact, this has been an active area of research recently. Examples of existing works that look at the dynamics of users to AP association includes [56-95], association durations at APs [96 - 100], traffic at APs [101 -109], and events counts for APs [48, 49]. All these efforts are attempts to understand the usage patterns behavior of wireless users, APs, location, and 802.11 networks as a whole so that administrators can take timely and proactive management actions when certain 802.11 usage behavior is evident.

Generally speaking, 802.11 networks are never free from problems. At any given moment in time one or more access points and users in the 802.11 infrastructure are likely to encounter problems such as overload, interferences, intermittent connections or authentication failure, and in some cases lack of coverage. To quickly and efficiently detect problems in these infrastructures requires development of automatic diagnostic tools and models.

2.4 Network Management

2.4.1 Network Management Architecture

Any network, regardless of whether it is small or large, can benefit from some form of network management. For a small network, network management may involve simply monitoring of network activity using a protocol analyzer. As the scale of the network increases, this form of network management proves to be expensive and labor-intensive. Thus, for a large-scale network usually network management involves the use of a distributed database, polling of network devices, and high-end workstations for displaying graphical views related to changes in network topology, usage, and traffic [15]. Taken as whole, network management is a service that uses different tools, applications, and devices in order to help human network administrators in monitoring and maintaining proper functioning of the network.

Despite the fact that each network management architecture is based on different building blocks, there exist certain aspects that are common to all network management architecture. Usually, a network consists of a number of different managed devices e.g. routers, bridges, switches, servers, and clients such as APs. Network management encompasses all activities involved in monitoring and maintaining these devices including altering of the configuration settings [16]. Usually, within these managed devices resides a certain type of **agent**. The task of an agent is to provide management information about the managed device, and also to receive instructions for the configuration of the device. It is also possible for an agent to reside outside managed devices; this type of agent is called a **proxy agent**, see figure 4.

On the backend there exists a **network management station** which normally offers a text or graphical view of the network or one of its managed components. Management station has the ability to provide this view through the use of a manager/ management application [16]. Typically, the exchange of information between the manager and agent is realized through a network management protocol. The exchanged information helps a network administrator to track usage and configure a network in response to the observed performance and detected faults. More than one network management station can exist in the network, each providing different views of the same part of the network or of different parts of the network.

It is possible for a network management system to operate in a centralized manner and in a distributed manner. In a centralized scheme, usually one computer system with a database handles almost all applications required for network management. In a distributed scheme, several network management systems can run simultaneously, each managing a specific portion of the network. A hierarchical network management system uses a centralized system at the root, with running distributed peers as children of the root [15, 16].

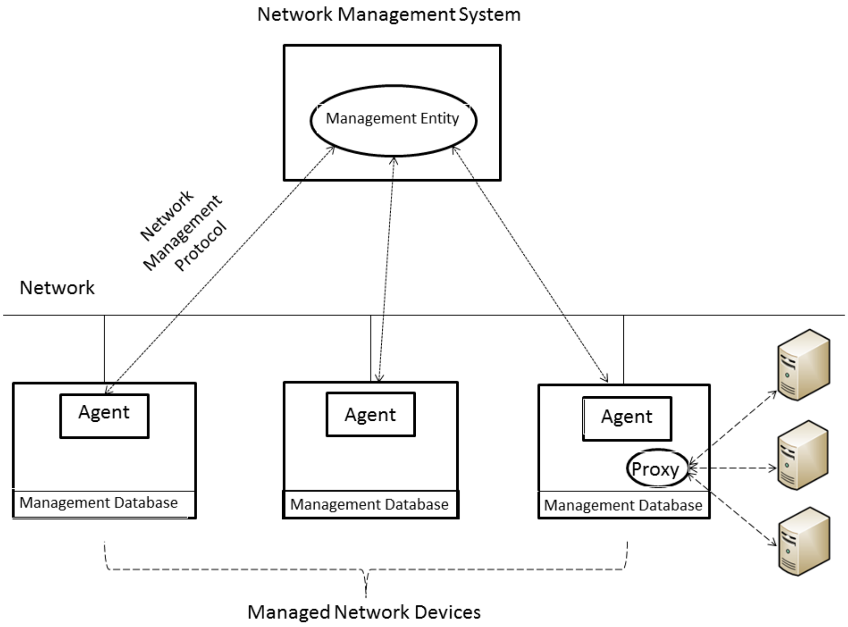


Figure 4. Network management system's architecture [15]

2.4.2 Simple Network Management Protocol (SNMP)

The Internet Activities Board (IAB) in the early days of Internet recognized the need for the development of a management framework to manage TCP/IP implementations [17]. The proposed framework consists of three components: 1) **Structure of Management Information (SMI),** which is a conceptual framework that depicts the rules for describing management information in the managed network system, 2) **Management Information Base (MIB),** which is a virtual database for keeping information about managed devices, and 3) **Simple Network Management Protocol (SNMP),** which is a protocol for allowing exchange of information between a manager and an agent within a managed network.

Essentially, the data that is supported by SNMP must abide by the rules related to the MIB objects, and must be defined according to the SMI. SNMP is an application layer protocol employed to retrieve and write variables in an agent's MIB within the managed systems. SNMP is based on an asynchronous request-response protocol further enhanced by trap-directed polling capability [17]. It is asynchronous in the sense that the protocol can send several messages without needing to wait for a response. Trap-directed polling allows for a manager to poll in response to a trap message initiated by an agent, particularly if an exception or threshold of some monitored variables has been reached [17]. Typically, SNMP uses UDP and operates in a connectionless mode. In addition to simplifying SNMP's implementation, this also provides the ability for a management application to communicate with many agents. For a convenience in this chapter, we use SNMP to refer to all versions of SNMP.

A typical manager and agent interaction using SNMP will proceed as follows. The manager will first issue gets request containing a unique request-id for matching the response to its respective request, a zero-valued error status, and other variable buildings. After receiving the request, the agent will reply with a response that contains the same request-id, a zero-values error status in case if there is no error, and the same variable buildings. If an exception occurs in some variables, certain error status will be returned. For example, if it happens the agent does not implement a particular variable, the `noSuchName` error status is returned, `tooBig` when response is too large to send, `badValue` when an invalid values or syntax is specified, and `readOnly` when a manager write a read only variable [17].

Remote Network Monitoring (RMON): RMON uses a technique named remote management to obtain monitoring data [18]. Using this approach, a network monitor also known as a probe is used to collect data from the managed device. The probe can be stand-alone or embedded within the managed device. Rather than communicating directly to the device, management systems use SNMP to communicate with an RMON agent in the probe [18]. This makes it easier the sharing of information among multiple stations. In addition, if it happens that management application loses connection to the RMON agent data can be easily retrieved. Because the RMON agent has ability to collect data even though connection to the management system is absent. Usually a probe has considerable resources, hence can easily keep historical statistical information that could be played back afterwards by the network management station.

2.4.3 Functions of Network Management Systems

The functions performed by a network management system can be categorized into the following five broad areas as far as FCAPS model is concerned.

Fault Management: Its main goal is to detect, isolate, notify, and correct faults encountered in the network [19]. The ability to detect problems quickly and efficiently in any network is critical because faults in the network can result into serious network service degradation. As such, fault management is one of the most implemented among the ISO network management elements. Typically, fault detection is accomplished by listening to alarms generated in real-time or through analysis of error log files. This is possible because most monitored devices are configured to send a notification when they encounter a fault, usually via Simple Network Management Protocol (SNMP). Only after a fault has been identified and analyzed can remedial actions be taken by administrators. These may involve actions such as debugging, rebooting, or replacing failed devices e.g. APs, routers, servers, and switches. This type of fault management is referred to as passive; the fault management system only gets to know of a fault after it has received, for example, an SNMP alarm [20]. One significant drawback with this approach is that the devices need not only to be intelligent enough to send alarms, but they should be in the appropriate state too: some faults can be so severe (e.g. AP crash) rendering the device unable to issue alarm, thereby allowing faults to go undetected. On the other hand, active fault management involves sending periodic requests, e.g. Ping or Traceroute, SNMP polling to monitored devices in the network, specifically for checking their accessibility and status and the value of certain variables [20]. If no response is forthcoming or a fault is diagnosed, alarm is issued. The main limitations of this method are three: first, it wastes considerable amount of network bandwidth due to the overhead involved in sending and receiving message in the network. Second, there is a chance of missing the actual thresholds, as every decision depends on the polling interval. Third, human administrators are still responsible for manually analyzing and interpreting the collected data to identify the causes of anomalies.

Configuration Management: Its main goal is to monitor and manage network system configuration information, including versions of different hardware and software elements on the network [19]. This area is equally important, as many network issues result from changes in configuration files, installed modules, and change of hardware models and software versions. Configuration information is usually collected and kept in an inventory database, which makes traceability of network devices configuration information easier [21]. In this manner, the updated configuration information can be easily collected by using e.g. SNMP or scripting. In addition, configuration management is useful in making large change because manually updating individual devices is a tedious, time-consuming, and error-prone task. Using an automatic configuration management system one specific change/update can automatically be applied to all devices in the network. This not only saves time, but also ensures consistency of configuration information among different network devices, which may in turn reduce possibility of errors. Owing to the heterogeneous nature of network devices, it is very important to implement a common interface that provides support for configuring all devices in the networks.

Accounting Management: Collects usage statistics for different users or departments so that bills can be generated and usage quotas can be enforced [19]. For non-billed networks ‘accounting’ is replaced by administration. In this respect, network administration aims to administer users of the network by granting passwords and network access permissions, also to administer the operations of equipment e.g. software backup and synchronization [22]. Typically, usage requirements for different organization differ significantly, which defeats the need of having a single accounting protocol. Nonetheless, one protocol commonly used for accounting is RADIUS protocol [35, 36]. RADIUS is an application layer client/server protocol that provides three main functions: 1) to provide authentication services for users and their devices so that they can access the network; 2) to grant authorization to authenticated users/devices to use the network and network services; and 3) to provide accounting for their usage of network and network services. For example, when a user is authenticated and connected, the Acct-Status-Type attribute will register this as “START” indicating that the request is the beginning of user service and when the user connection ends the Acct-Status-Type attribute will register this as “STOP”. The STOP records in the RADIUS server contains all the information in the start record plus additional usage information, such as access time, number of input and output bytes as well as number of input and output packets. For a usage-based accounting, this information can be useful for billing purposes and invoice generation [22]. The accounting data can be used to extract knowledge of overall network usage and for trend analysis. Trend analysis involves forecasting future usage, which is usually achieved through the use of statistical sampling techniques and models [22]. In fact, trending is equally important even for billed networks, because prediction of future usage helps administrators to prepare network and network resources to cater for any anticipated usage.

Performance Management: Ensures performance of the network and network services remains at acceptable thresholds [19]. Performance management may include activities such as monitoring, planning, and fine-tuning the network to fulfil performance requirements of the organization, specifically in terms of offered capacity, reliability, and latency [19]. Most performance problems in the network are related to capacity, for example the offered throughput of the network can be drastically lowered as a result of too many users sharing the same network element such as APs or network service such as streaming server. Performance information can be collected passively or actively through management systems that implement SNMP, and configured to alert network administrators when performance indicators move above or below given thresholds [15]. By collecting and analyzing performance data over time, knowledge of usage baselines and trends for different network elements and services can be established. This may allow network administrators to plan and perform network maintenance before a given capacity problem can cause network down time [23]. The major challenge of baselining and trending approaches is on dealing and processing huge amount of information such as the one generated by a large-scale network. Therefore, it is of utmost importance to define clearly the variables to monitor from the collected usage data and the threshold values that require action.

Security Management: This network management function is responsible for protecting network and network resources against unauthorized access. It is also responsible for managing user rights and access privileges so that only legitimate users have access to appropriate network resources [19]. Specialized analysis tools are often used to identify assets and their respective threats, and also to rate overall network system vulnerabilities which in turn makes mitigation of any potential security problems easier [24]. Appropriate security measures need to be configured on the network and network resources in order to ensure that sensitive data cannot be accessed or changed e.g. via SNMP or rogue AP. Additionally, key management is also critical to the security of a network and network resources. Rules must be imposed regarding the use of passwords. In this respect, all passwords used must abide by the organization standards and guidelines such as minimum string length, and mixture of characters and numbers to make a guessing of password difficult. The starting point for any good security management implementation is on the organization security policies and procedures. These are out of scope of this thesis, and are not further addressed.

2.5 Trace-Based Analysis as a complement to Network Management

2.5.1 SNMP-based Network Management

Many network management efforts make use of protocols such as the simple network management (SNMP) protocol and data models (e.g. management information base (MIB)) for automating the task of collecting information from network devices and for communicating back to network administrators [25][26][27]. With these approaches, network administrator manually analyzes and interprets the collected data before attempting to carry out any service update, configuration, or network upgrade. Network management following these approaches is highly influenced by the skills of the network administrator.

2.5.2 Web-based Network Management

Web-based approaches and common object request broker architecture (CORBA) are among other network management efforts [28][29]. These approaches use web services as a management framework in simplifying network management operations e.g. configure device, turn it off and on, and collect alarms. The main limitation with these approaches is that, even a simple task such as introducing a new network variable or defining new MIB entry requires the overhead of taking the whole managed system down [30]. Again, the operation of the managed networks is constrained by the expertise of human network administrator.

2.5.3 Automated Network Management

The increased complexity of managed network systems led to development of management tools that try to alleviate the task of network management from human network administrators. Examples of these approaches include mobile agents, active networks, and policy languages [31][32][33]. A major limitation of these automated management approaches is their reactive nature. In the sense that, all management decisions depend on the current system diagnosis, leaving no room for past observation or future prediction [34]. As such, these approaches are unable to adequately evolve and adapt with changing organization objectives and usage demands.

2.5.4 Trace-based Network Management

Despite the network management approaches mentioned above, trace-based management approaches appear to provide a promising solution to most of the aforementioned problems. To achieve large-scale data collection, research community uses existing 802.11 infrastructures to obtain 802.11 traces. Typically, 802.11 access points (APs) and servers (e.g. RADIUS) in the 802.11 networks are configured to monitor the association and usage of individual 802.11 users. In this manner, relatively rich data set related to usage characteristics of APs and users that spans seconds, minutes, hours, days, months, and years can be maintained. In addition, the research community has been organizing traces over the years and several websites for archiving the relevant traces have been maintained. Two prominent examples of such websites are [37] and [38]. These websites grant accessibility of 802.11 traces and other relevant resources to the research community. Example of research works that use these traces can be found in the following references: user mobility modeling [53, 55, 57, 60, 61, 65], user registration pattern [89, 91, 92, 93], access duration and traffic [96, 104, 105].

Trace-based Analysis and Network Usage: Systematic analysis of the collected 802.11 usage traces can help to understand deeply usage dynamics of different components within the managed network. This can be accomplished for example, by defining a set of models based on the collected underlying data that characterize the usage patterns of the access points, users of the network, applications running on top of the underlying infrastructure, locations of the environment encompassing the network, and the overall aggregate infrastructure usage properties. For example, in the previous research large-scale trace analysis has proven helpful to unearth hidden trends in 802.11 usage such as user mobility [56 - 80], access point popularity and their periodic usage patterns behavior [51 - 64], also preferences of applications by users and popularity of their devices [100 - 116]. Other interesting findings from trace analysis include: time-varying Poisson processes for modeling users' arrival to APs [58, 63, 89, 90], power-law distributions for modeling distinct user groups with commonalities [82, 83], Bi-Pareto distribution for modeling encounter events between individual nodes [81], and Weibull regression model for approximating the arrivals of traffic flows at individual APs [107].

Trace-based Analysis and Anomaly Detection: Systematic analysis of the collected 802.1 usage data can help to detect faults or anomalies in the components of the managed network. Example of useful findings in the literature related to detection of anomalies based on the analysis of 802.11 traces include: detection of repeated handoffs in client connections at the same AP during overload [117, 118, 119, 155], arrival of flash crowds at APs [123], detection of unusual events or happenings in physical space [39], and detection of other anomalous usage patterns such as short sessions [40]. In addition to detection of anomalies, statistical models for characterizing occurrences of different faults can be established. These models enable estimation of occurrences of faults at some stage during network operations (e.g. AP failure or AP overload, intermittent connections at AP etc.), allowing network administrators to take precautionary measures, especially in reducing the impact of the network anomalies or even preventing the anomaly from occurring at all.

Trace-based Analysis in other Domains: Other possibilities exploited to collect usage traces include cellphone traces [41 - 43]. Though cellphones are perhaps the most popular wireless devices used, large-scale traces are difficult to obtain from the cellular phone operators owing to privacy concerns. There are also emerging efforts in collecting vehicle movement traces preferably through GPS positioning system, for example [44]. However, this effort requires active reporting from the monitored vehicles, hence does not scale well. A study of human physical mobility has relatively received greater attention in the literature. For example, authors in [45, 46] tried to understand human global mobility patterns by examining human travel behavior through bags and money circulation properties. These efforts indicate importance of trace-based analysis for different empirical studies.

Summary: Management techniques developed based on the collected usage traces are likely to be proactive rather than reactive, in the sense that management decisions can be made based on analysis of information related to present usage state of the network or managed components as well as based on the knowledge of the predicted future usage state of the network. For example, through our probabilistic models for AP usage (chapter 4), future usage pattern of the AP can be depicted based on either present day usage, or based on specific number of consecutive previous days or similar past days usage patterns behavior. In this manner, baselines for usage of APs on different days can be established which can later help to identify APs and locations surpassing planned capacity. In addition, models or algorithms developed based on the collected 802.11 usage data can be fine-tuned with little effort to suit different 802.11 environments. In this case, what one needs to do is just to change parameters of the models to include parameters of the new underlying 802.11 environment. As a side note, to generalize findings beyond a specific environment, it is important to consider analyzing multiple traces from different networks using the same exact method.

2.6 The need for A Framework

It was argued in chapter 1 that, as usage and scale of 802.11 networks grows, so grow the challenges of managing these networks. One argument is that maintaining baseline knowledge of infrastructure usage properties, including usage patterns of individual APs, users, and locations, and their likelihood of encountering problems by administrators becomes increasingly difficult due to scale. On the other hand, due to the unreliable nature of the wireless medium, 802.11 users are likely to suffer performance and connectivity problems while using these 802.11 large-scale infrastructures, e.g. QoS degradation of their traffic, intermittent connectivity of their connections, and in some cases authentication failure. These problems can be caused by AP interference, AP overload, crashed AP, or weak RF signals due to RF holes/dead spots. To detect problems by 802.11 administrators physically visiting the place after receiving a phone call, e.g. from troubled 802.11 user, can be impractical due to size but also it is possible that when administrators arrive to the spot the problem may have disappeared only to reappear later on.

To complement existing management techniques, in this thesis we propose a framework for usage modeling and anomaly for the large-scale 802.11 networks, based on the collected 802.11 usage data. Since there is no any known network management solution that allows network administrators to perform both performance and fault management using an easy access data set, our proposed framework use 802.11 AP session data collected centrally at RADIUS authentication server. This makes easier accessibility of user-AP association records for all APs on different parts of the large-scale network. With this readily available data set, extensive analysis pertaining to the underlying usage properties of 802.11 APs and their associated problems can be easily accomplished.

In connection to this, our proposed framework is based on this collected 802.11 AP usage data and consists of two research thrusts, namely usage modeling and anomaly detection (see figure 1), which are captured by two network management functions: performance and fault management, respectively. Combining these two network management functions allows network managers that use our proposed framework to prevent problems in the usage of the large-scale 802.11 networks before they affect network and network service. For example, through the usage modeling feature, network administrators will be able to understand and predict usage in the infrastructure, thereby allocating network resources accordingly. In this manner, in addition to establishing usage baselines for APs, capacity problems of large-scale infrastructures can be prevented before affecting network services. For example, if there is knowledge that usage is going to exceed for some APs in a given day, administrator may decide to temporarily install additional APs in the vicinity of those APs so as to redistribute the load. On the other hand, the anomaly detection feature of our framework will help to deal with network connectivity issues. This will help in detecting connectivity problems quickly and efficiently, giving network administrators the opportunity to deploy counter active measures, for example to change the configuration parameters of APs, reboot, add, relocate APs, or remove problematic APs. And this could be done before users start calling to complain about their wireless network usage experience.

To develop our framework, we started by looking at the extensive 802.11 data sets collected from our campus 802.11 network and identified the important characteristics of data that we intend to work upon. Details of these data sets are provided in the subsequent chapters of this thesis. Following these observations, we engage first into a modeling effort of AP usage in chapter 4, where we train and evaluate different time-independent and time-dependent models for characterizing AP usage, based on AP keep-alive event counts. Then in chapter 5 we focus on anomaly detection of AP usage and we investigate a method for detecting abrupt ending of AP connections from massive collected 802.11 usage data. We propose statistical models for characterizing their occurrences with respect to aggregate 802.11 network usage. In chapter 6 we present an algorithm for characterizing different anomaly-related patterns of AP usage resulting from the occurrences of abrupt endings. Therefore, patterns such as AP interference, AP overload, AP crash, user authentication failure, and intermittent connectivity were detected, characterized, and modeled. Moreover, we map all the contributions of our work to network management FCAPS model, specifically performance and fault management, in order to support the importance and usefulness of the observations.

Chapter 3

Related Work

3.1 Introduction

In this chapter we analyze existing works on 802.11 usage modeling and anomaly detection. This is because our framework proposed in this thesis includes these two features in aiding both performance and fault management of large-scale 802.11 networks. We begin with the analysis of the early trace-based research on 802.11 usage until recent works with an emphasis on APs, users, and traffic modeling. We analyze works conducted on the traces of university campuses, corporate organizations, and public 802.11 infrastructures. In addition to early works about general statistics subsection 3.2.1, we present analysis of the recent works related to user mobility modeling at AP in subsection 3.2.2, user encounters patterns in subsection 3.2.3, user registration patterns in subsection 3.2.4, user access durations in subsection 3.2.5, and traffic characterization in subsection 3.2.6. From AP perspective, all the above mentioned items constitute to the aggregate usage of the 802.11 AP. When considering the problem of modeling 802.11 AP usage to aid performance management of a large-scale 802.11 network, it is increasingly important to analyze all aspects of user activities that may amount to overall usage at the 802.11 AP. We also include the analysis of public 802.11 infrastructure usage in subsection 3.2.7.

Moreover, for anomaly detection in 802.11 networks, we analyze relevant existing works not just limited to the collected 802.11 usage data, but simulations and testbed works are also included. This is due to rarity of works in the literature that tried to detect anomalies from the collected 802.11 usage data. The main emphasis is placed on the detection of connectivity anomalies in the usage of 802.11 access points. We focus on the features and techniques used for the detection of AP overload in subsection 3.3.1, AP halt/crash in subsection 3.3.2, AP interference in subsection 3.3.3, rogue AP in subsection 3.3.4, and other performance anomaly of 802.11 networks such as capture effect, repeated handoffs, and MAC misbehavior in subsection 3.3.5. The first three items mentioned above are in particular related to our anomaly detection work of chapter 5 and 6. Also, in addition to connectivity problems caused by rogue AP and various performance anomalies of 802.11 networks, these three anomalies (AP overload, AP halt/crash, and AP interference) are responsible for majority of connectivity problems in these large-scale 802.11 infrastructures. We will review these works in this chapter.

3.2 Usage Modeling

3.2.1 General Statistics

In the initial stage of 802.11 trace-based works, most researchers focused on understanding how wireless users generally use the deployed 802.11 infrastructures. In many of these works, basic statistics about user behavior and network performance were collected and analyzed. For example, Tang and Baker [51] collected a twelve-week 802.11 trace from a university campus, aiming to characterize global properties of 802.11 infrastructure usage in the academic environment. In their analysis, they observe most users in campus are stationary and utilize the network only for web-surfing, session-oriented, and chat-oriented activities. Additionally, they find that the peak throughput is usually caused by a single application (HTTP) and that in overall incoming traffic dominates outgoing traffic. Later, Tang and Baker [52] analyzed network traces of a metropolitan-area wireless network, aiming to understand user behavior in a typical daily life situation. In their analysis, they find the network was used mostly during the day and evening hours, and on average users associate with few APs which are in close proximity geographically. Along the same direction of research, authors in [53] tried to characterize user behavior and network performance from a three day conference network trace. In their analysis, they observe that users are evenly distributed across all APs. In addition, load distribution between APs is highly uneven and web traffic is responsible for 46% of the total bandwidth among all application traffic mix.

In [54], authors extended the work of Tang and Baker [51], by considering larger traces covering almost all buildings on a campus, including residential buildings in analyzing user behavior and network traffic. They find that residential traffic is the most dominant one among all other traffic, and web protocols account for (53%) of the total traffic volume. In [55] authors analyzed over seventeen weeks of trace data from a mature 802.11 network that includes about 550 access points and 7000 users. They compared the findings from this trace to the trace collected two years ago after the initial deployment of their WLAN [54]. In their analysis, they observe substantial change in applications used in 802.11 infrastructure including significant increases in peer-to-peer, streaming multimedia, and voice over IP (VoIP) traffic, with on-campus traffic exceeding off-campus traffic. This is contrary to the initial stage where web and off-campus traffics were dominant components.

Although the analyses presented by these works help to gain general understanding of user behaviors and network performance, most of these works do not try to model 802.11 usage. To improve upon this shortcoming, most recent 802.11 trace-based works attempted to focus on modeling user behaviors; one particular important aspect is the mobility of 802.11 users.

3.2.2 Mobility Modeling

User Mobility: There are numerous efforts in the literature that tried to exploit the ubiquity provided by 802.11 networks to understand and model mobility of users among available access points (APs), buildings, and across the whole 802.11 infrastructure. For example, authors in [56] modeled user mobility based on how frequently users visit various APs together with their duration of stay in those APs. They find that users spend a large fraction of their time at a single home AP, whereas the probability distributions of their movement and stay time follow power laws. Authors in [57] considered the number of unique users associated to APs during each hour, aiming to classify user mobility patterns on hourly basis. The analysis of a one-year 802.11 trace indicates both user mobility and AP popularity depend on the academic calendar and their periodic behavior depends on the proximity to other APs. Authors in [58] modeled influx and outflux of user movement at AP from a two-month period Dartmouth traces. By counting the hourly number of users at each AP and through the use of Discrete Fourier Transform (DFT), they observe repeated temporal pattern every 24 hours at each AP, by considering the number of associations of the same hour in different days they then aggregate multiple days' association vectors to form a single 24-elemet vector for each AP and cluster APs based on their peak hour of visits. They find that average users' arrival rate and the distribution of the daily arrivals for each cluster follow non-homogeneous Poisson processes. Authors in [59] studied individuality of users' mobility patterns from eight months Dartmouth traces. Without using geographical information, they proposed a mobility clustering aware algorithm based on graph theory, which uses the number of roaming events between APs as a metric for estimating proximity between APs. Using their algorithm, they were able to depict individuals' mobility as well as to categorize the resulting clusters into places with social meaning and paths.

Focusing on mobility patterns of hand-held devices rather than laptops, authors in [60] studied a method of estimating the user's physical location from 802.11 traces and of mapping the coordinates of the campus and GPS data for each AP. By observing association events of mobile devices such VoIP devices and PDA on each workday, they find that pause time and speed distributions each follow a log-normal distribution, and the direction of movements follow the direction of popular roads and walkways on the campus. Along the same direction, authors in [61] analyzed the mobility patterns of PDA users in a campus WLAN based on their proximity when they are within communication range of each other. Authors observe their proposed evolutionary topology model results in a completely connected graph only when users' moves within the indoor locations of campus, and once these users moves throughout (outdoor) campus they create numerous islands of disconnected graphs, which can be captured by their other proposed campus waypoint model.

Other interesting research on user mobility include [62], where authors analyzed 802.11 traces collected from three university campus WLANs, aiming to compare user association behavior to access points in these environments. By looking at user association patterns within each AP and between APs, particularly at the repetitive nature of these associations, they find that a significant portion of users' exhibit off-on usage pattern, that the majority of users visit only a small subset of APs, and that most users show periodic pattern of visiting the same APs. In [63], authors developed a simple mixed queuing network model of user mobility among available access points in a campus network. They consider list of APs as a mixed network of infinite server queues and model the exogenous arrivals of users to AP as a Poisson process. From their experimental analysis, authors observe the aggregate user movements to the campus network followed a Poisson process.

Focusing on buildings rather than APs, authors in [64] analyzed a year of trace data collected on the ETHZ campus, aiming to characterize the movement pattern of wireless users in campus buildings as well as to predict their future locations. Their experimental results indicate that user regularly visit a list of buildings with some probability, and that the suitable choice to model the binary user-building visitation vectors is a Mixture of Bernoulli's distribution.

The user mobility characterizations of all the aforementioned campus WLANs yielded almost similar results. The usage is generally diurnal in nature and only a small fraction of devices are mobile. This can be attributed to the fact that each of these environments are of the same nature with a user base that involves in similar activities. Most of these works consider distribution of association events from few users rather all users associated to an AP, to depict mobility models for individual user or group of users. Therefore, the resulting models can hardly capture the overall usage characteristics of 802.11 APs in large-scale networks.

Location Prediction: There exist works in the literature that tried to predict the AP that users will associate to in the future based on their previous mobility patterns. For example, authors in [65] presented results on user mobility from the empirical evaluation of location predictors using four major families of domain-independent predictors i.e. Markov-based, Compression based, Prediction by Partial Matching (PPM) and Statistical Parametric Matching (SPM) predictors. Based on the transitions between APs, they find that low-order Markov predictors performed better in predicting future locations than the more complex and more space-consuming compression-based predictors. Similarly, authors in [66] investigated a series of prediction methods for predicting users' mobility, using Markov chain family of predictors (ranging from order 1 through 3) based on the number of distinct and total APs a user has visited. They find that building level prediction has significantly better prediction accuracy than AP level prediction, and also, the more mobile the 802.11 users become the larger the gap between building and AP level predictions. In [67], authors built a mathematical model for characterizing steady state and transient behaviors of user mobility in WLANs. By looking at the steady-state distributions of semi-Markov models at different time scales and by characterizing user-AP associations based on their correlation in time and location, they propose a timed location algorithm for predicting the future locations of users and their association duration. In [68], authors examined the regularity and predictability of human mobility patterns. They collect user's mobility every two minutes using GSM, WiFi, and GPS from 10 users over a period of two months. Their experimental results indicate that a location-dependent predictor is better than a location-independent predictor for predicting temporal behavior of the individual user, and furthermore, that the duration of stay at a location is strongly correlated to the arrival time at the current location and the return-tendency to the next location, rather than recent location sequences.

Along the same direction of research, authors in [69] presented an approach to characterize steady-state and transient-state user mobility behaviors based on the Hidden Markov Model (HMM). They assume that each node is able to periodically locate itself via the low-cost GPS devices. For example, if a node locates itself at a period T, then its velocity vector is captured as HMM state and a vector displacement in this case is considered as mobility. From observed HMM mobility states they were able to depict the steady-state movement i.e. the cumulative moving direction of a node during a period of time, as well as to predict transient-state node mobility behaviors in the near future. In [70], authors applied Markov renewal processes (MPR) to model user mobility and for predicting the likelihoods of the next AP transition, together with the expected duration between the transitions for an arbitrary 802.11 user. They observe that their proposed MRP-based mobility prediction achieved significant improvement, in terms of prediction accuracy than what could be achieved by using transition probabilities alone.

However, closer look at the results of these works reveal that there remain a significant number of users with poor prediction accuracies, specifically if the movement of a node is very random. This can be largely attributed due to the Ping-Pong sessions rather than actual user movement. Furthermore, most of the proposed predictors failed to make a prediction when the recent context, e.g. a sequence of APs, has not been previously seen.

Ping-pong effect and Mobility: There are also efforts in the literature that tried to remove users-AP association patterns which are usually confused as mobility patterns, while indeed they are not. Typically, in a densely deployed 802.11 network mobile stations continuously tends to alternate association between access points in searching for better association. Under this greedy choice of association, it is possible for some user sessions to appear in different APs which are further apart, even without users' physically moving into them. For example, authors in [71, 72] identified these elusive patterns that are normally considered to be user movements in constructing empirical mobility models. Their analysis reveals that about 38–90% of transitions are not connected to actual user movements. In [72], authors analyzed 802.11 trace data in comparison to SNMP data, and show that by removing these elusive registration patterns, the fidelity of empirical mobility models can be significantly improved.

In fact, this is in agreement to our work, since the variables we consider in this thesis are daily counts of keep-alive events generated every 15 minutes for refreshing mobile stations association to AP. Ping-Pong sessions are usually sessions of very short duration, few seconds or minutes, and are unlikely to last 15 minutes. This eliminates completely the effect posed by the Ping-Pong sessions in building our AP usage models.

Human Physical Mobility: There are further efforts in the literature that try to capture physical movement of wireless users, aiming in particular to avoid unrealistic assumptions of the existing mobility models used in network simulators. Most of the mobility models used in the existing simulation tools such as ns-2 are limited to the concept of independently and identically distributed random variables (i.i.d). The main assumption with these models is that every node behaves statistically the same throughout the entire simulation period: choosing randomly the next location, speed, and direction. For example, authors in [73] proposed a Semi-Markov Smooth (SMS) mobility model that characterizes the smooth movement of mobile users to eliminate sharp turns, abrupt speed change, and sudden stops exhibited by existing mobility models used in network simulations such as the Random Waypoint (RWP) mobility model [180]. Their experimental results indicate that the relative speed between a pair of mobile nodes follows a Rayleigh distribution. They also find that the CDF of the link lifetime and the average node degree indicate that the network connectivity based on the RWP model can be somehow over optimistic. Authors in [74] propose an integrated quality of service (QoS) aware mobility and user behavior model. The proposed mobility model implements weighted random waypoint (RWP) mobility, where the next waypoint is chosen based on the correlation coefficient between the number of registered users and the observed load at the AP, rather than choosing next way randomly. In [75], authors presented a mobility model called SLAW (Self-similar Least Action Walk) containing features such as truncated power-law distributions of flights, pause-times and inter-contact times, fractal way-points, and heterogeneously defined areas of individual mobility. Experimental results indicate the proposed SLAW can effectively extracts mobility patterns arising from people with some common interests like students in the same campus or gathering places such as street malls and restaurants that most people tend to visit during their daily schedules. In the same direction of research, authors in [76] find out human walks at outdoor environments less than 10 km exhibits features statistically similar to Levy walks. These features include heavy-tail flight and pause-time distributions, and the super-diffusion followed by sub-diffusion in a confined area. Authors in [77] investigated the properties of (Small Word in Motion) SWIM model, in particular how SWIM is able to generate social behavior among the mobile nodes and how SWIM is able to model user mobility with a power-law exponential decay dichotomy of inter-contact time and with complex sub-structures (communities), similar to the ones observed in the real data traces. They simulate three scenarios and compare the synthetic data with trace data in terms of inter-contact, contact duration, number of contacts, and presence and structure of communities among nodes; they find a significant matching and SWIM was able to extrapolate the above mentioned properties of human mobility behavior. Focusing on actual user movement between physical locations, authors in [78] attempted to analyze the short and long term movements between places. They measure the degree of connectivity between two places, by counting the number of movements observed between two places within a given time period, and observe strong connections between movements in nearby places e.g. places located in the same building or nearby buildings. Also, they observe that short duration movements are more frequent than long duration movements, and that the time span between movements can be well approximated by a power law distribution.

There exist also efforts in the literature that tried to understand mobility patterns of moving vehicles rather than human movement, particularly in relation to the quality of their connections as they pass by and connect to nearby base stations (802.11 APs). For example, in [79] authors find that the interval between a vehicle coming into and going out of range of a base station is often impaired by intermittent periods of very poor connectivity. Their experimental analysis indicate that knowledge of past connectivity and movement path can be used to predict regions where poor connectivity are more likely to occur as well as regions where the vehicle is likely to experience good connectivity. Also, authors in [80] conducted a study in an attempt to improve connection quality available to vehicular WiFi clients based on measurements from testbeds in two different cities. They find that current WiFi handoff methods where clients communicate only with one base station at a time lead to frequent disruptions in connectivity. Motivated by this, they develop a protocol (ViFi) that opportunistically exploits base station diversity to minimize connectivity disruptions of moving vehicles.

Again, most of these works consider association events of only few specific individuals on a small set of APs, rather than association events of all users from all APs. In addition, for most of these works their evaluation is based on a simulation, testbed, and a small scale network. Therefore, applicability of these proposed approaches to a network with high number of users, APs, and locations is not guaranteed. In addition to user mobility, there are further attempts along the same line of research, aiming to understand encounter patterns of individual users and group of users from the association patterns of APs. We discuss these works in the next subsection.

3.2.3 User Encounter Patterns

There are several works in the literature that attempt to establish inter-relationship and similarities between individual users or between groups of users by observing their association patterns when using APs and buildings across 802.11 infrastructures. For example, authors in [81] examined mobile node encounter patterns from campus and enterprise wireless networks using a graph analysis approach. They look at the asymmetry of inter user relationships (node pairs) based on time, duration, and frequency a given user pair spends together at APs. They observe the encounter events for each mobile node followed the Bi-Pareto distribution, which suggests the behavior of mobile nodes is not independent and identically distributed (iid); whereas the empirical distributions of the friendships indexes followed the exponential distribution, with few pairs showing strong relationships. Authors in [82, 83] attempted to analyze users behavioral patterns from 94 days traces collected from the Dartmouth campus. They generate a daily association vector based on individual user-AP associations and use average minimum vector distance (AMVD) to classify the entire population into several user behavioral groups. They observe hundreds of distinct user groups with commonalities in their usage patterns which followed power-law distribution.

Additionally, they employ their classification methods to different traces, e.g. from the University of Southern California (USC), and observe repetitive behavioral trends and user grouping, indicating consistency of user behavior between the two universities.

Other extended research in this area include [84], where authors proposed a time-variant community (TVC) mobility model, not just for depicting locations (APs) that are visited by the nodes so often but rather for estimating periodical re-appearance of nodes at the same location. They compute meeting probability between two nodes in a unit-time slice and the meeting probability of each time period including the time of the last period in which the meeting event between nodes occurs, and propose to use a time-variant Markov chain to capture the spatial and temporal dependencies in nodal mobility. They find visiting preferences of users to various locations followed power-law distribution, while the probability of nodes re-appearing at the same AP is higher if the time gap is an integer multiple of days, specifically 7 days. Authors in [85] extended TVC model in [84] by analyzing vehicular and human encounter traces, in addition to WLAN traces. They find location visiting preferences and periodical reappearance are also prominent mobility properties in vehicle mobility traces, and also the movement is similar to that of WLAN users. Furthermore, authors in [85] analyzed log encounters between nodes carried by participants as they move around the premises at the INFOCOM conference; they show that TVC model is generic enough to mimic the encounter properties of mobile human networks in terms of inter-meeting time and the encounter duration distributions.

Along the same direction of research, authors in [86] presented two clustering algorithms based on K-Means, aiming to determine the number of groups and the membership of mobile nodes within groups from association trace of mobile nodes. They evaluate their algorithms using both synthetically-generated traces with a known group structure and a real-world trace from a military scenario. Results show the number of groups and node membership can be accurately extracted by both algorithms, particularly when the number of groups is reasonably small as accuracy decreases when the number of groups becomes significantly larger. In [87], authors examined distribution of inter-contact times between mobile devices. They consider return time of a mobile device to its favorite location (AP) to establish inter-contact time between device pairs. They find that there is a specific characteristic time (order of half a day), beyond which the inter-contact distribution decays exponentially, however, before this value the distribution mostly followed a power-law. They also show that simple models such as random walk and random waypoint can exhibit the same dichotomy in the distribution. Authors in [88] studied regularity in weekly visiting patterns of users in three real-world datasets, specifically a metropolitan transport system, a university campus, and an online location-sharing service. In their experimental evaluation, authors identified a core group of individuals in each dataset that visited at least one location with near-perfect regularity. In addition, they observe a strong correlation between an individual’s most-visited location and regularity.

Most of these works also considered association events from only certain users or group of users. This means that a great number of users are left out in the analysis, making these models not suitable for overall capacity planning of large-scale networks.

In addition to user encounter patterns at APs, there are several attempts along the same line of research aiming to understand general user registration patterns at APs. We review these works in the next subsection.

3.2.4 User Registration Patterns

User association pattern has been one major focus in studies about 802.11 networks. The focus of most previous works in this area is on modeling user arrival and departure patterns to APs, and buildings across 802.11 infrastructures. For example, authors in [89] attempted to model the arrival of wireless clients at the access points (APs) of campus 802.11 network. They look at user-AP associations at different time intervals and cluster APs based on their visit arrival to APs and buildings. They find out Time-varying Poisson processes can be employed to model the arrival processes of clients to both APs and buildings. Authors in [90] studied several Wi-Fi hotspots deployed in a nationwide network for different types of venues, ranging from small coffee shops to large enterprises within two large cities in US. They observe the user arrival patterns to APs shows a significant difference between week days and weekends for all venues, even though the process of user arrivals at APs differs from venue to venue; they demonstrate it is possible to model these processes using the common model type of non-stationary Poisson processes. Other interesting work include [91], where authors modeled spatial registration patterns of user movement at campus APs from a two year trace collected at Dartmouth campus. Based on user-AP associations, they build a transition graph with each AP as vertex and transition between APs as edge. They clustered APs into disjoint clusters using Markov clustering tool and find that cluster size distributions are highly skewed with intra-cluster as well as inter-cluster transition probabilities followed Weibull distributions. Authors in [92] modeled the associations of a wireless client as a Markov-chain, in which a state corresponds to an AP that the client has associated with. Based on the history of the past transitions between APs, they build a Markov-chain model for each client, and use this model to predict next association that a client will establish as it is roaming the wireless infrastructure.

In the same direction of research, authors in [93] modeled temporal registration patterns of user associations by considering not only the number of user visits to an AP, but also the time spent at each AP before a user moves from one AP to another. They find that whenever there is a transition from a very popular AP to an unpopular AP, the residence times before the transition tend to be much larger than transitions from an unpopular AP to a popular AP or transitions between APs with similar popularity. Authors in [94] analyzed the regularity of the users in accessing the WLAN at different days from traces of two different campuses. They find out population accessing the networks is mostly composed by infrequent users. They also show that the distribution of the frequency of reconnections to the WLAN is not uniform, and users on a small campus are more likely to reappear on different days than on a large campus, where the population is more heterogeneous. Moreover, they show that usage behavior at the libraries of both campuses passed Chi-Square goodness of fit test, while the geometric distribution fitted the behavior in the building with classrooms. In [95], the same authors examined usage pattern of the library in the main campus of the Technical University of Catalonia (UPC). They find existence of daily and weekly repetitive patterns in WLAN usage, half of the population accesses the WLAN once during each month, and many users associate to only one of the twelve possible access points, indicating many users are static. The behavior on Fridays and the weekends were different than that of the first four days of the week and the proximity of final exams resulted in slight increases in the average amount of active devices registered to available APs.

These works analyzed total user associations only from a specific set of APs and buildings such as library. Therefore, little is known about applicability of these models to all APs, buildings, and locations within large-scale 802.11 networks. Furthermore, there is no consideration of Ping-Pong sessions in the construction of these models. In addition to user registration patterns at APs, there are also several attempts in the literature that aim to characterize user access durations at APs. These works are discussed in the following subsection.

3.2.5 User Access Durations

Characterization of user access durations has been researched extensively in the studies about 802.11 network usage. The focus of most previous works in this area is to understand time users normally stay associated to APs. For example, authors in [96] analyzed duration spent at each AP by considering continuous user session to AP without any disconnection (i.e. without re-association between consecutive associations). Their analysis show that, mobility and building type affect the session and visit durations at APs, also they observe as the mobility increases the visit duration tends to decrease stochastically, while opposite happens for session duration; hence proposed a family of Bi-Pareto distributions to model the associations and session duration at APs. Authors in [97] characterized usage of the campus network in terms of overall infrastructure usage, user mobility, sessions, and access durations. They observe the number of APs visited per user is influenced by geographic proximity of the locations, and in overall, the number of sessions per user followed a Logarithmic distribution while durations for stationary session followed a two-parameter Weibull distribution and mobile session durations followed an Inverse Gaussian distribution. Authors in [98] developed a wireless user model from analysis of five different traces. They define several users' states (i.e. active, idle, sleeping, and gone) and state transitions, and propose to employ a hidden Markov model to these state transition matrices. They find that the user models are similar across all five traces even though the traces were collected at different venues (library, coffee shops, and conference) and residing time to APs followed a generalized Pareto distribution.

In the same direction of research, there exist several works that try to understand user access time from simulation and testbed experiments. For example, authors in [99] studied the impact of mobility models on cell residence time for WLAN. They simulate different scenarios of AP density and observe that the average cell residence time decrease when a memoryless movement pattern is followed (e.g. Random Waypoint) and increase when smoother movement patterns are followed (e.g. Gauss-Markov); they further indicate the cell (AP) residence time can be better characterized by lognormal distributions. Author in [100] designed and implemented a general framework for behavior-aware dwell prediction with and without the aid of client sensor data. In their testbed experiment, mobile devices (smartphones) are programmed to periodically report their sensor readings to the AP (e.g. accelerometer and compass), a support vector machine (SVM) classifier accepts the readings and combine with measurement features from WiFi RSS, and therefore, based on training data from past data a classifier predicts users likely dwell time.

Most of these models consider aggregate association duration at AP regardless of the type of connections. Hence the accuracy of the resulting models can be highly influenced by the Ping-Pong connections at AP. In addition to user access duration at AP, there several works in the literature that attempted to characterize and model traffic. We review these works in the next subsection.

3.2.6 Traffic Characterization

Focusing on the infrastructure usage and network performance rather than user behavior, there exist a number of studies aiming at characterizing traffic of access points and aggregate traffic load of 802.11 infrastructure networks. For example, authors in [101] proposed a time-series forecasting method for characterizing traffic of APs. Using their proposed methodology, they observe the aggregate hourly traffic for all APs in the infrastructure exhibits diurnal and weekly periodicities, while similar trend is observed in the hourly traffic for several APs. Authors in [102] provided Singular Spectrum Analysis of the structure of traffic load measured in a large-scale campus-wide WLAN. Using their approach, they observe the time-series of traffic load at a given AP has a small intrinsic dimension, which can be modeled using a small number of leading principal components, while the residual components can be exploited to capture irregular variations (i.e. a stochastic noise). Authors in [103] proposed a traffic prediction mechanism using the Recursive Least Squares algorithm. Using their proposed method, they were able to predict future traffic load at APs in the time scale of few minutes, although prediction accuracy was constrained by the amount of history size used in the training process.

Other interesting works include [104], where authors performed a system-wide characterization of the workload of wireless access points (APs) in a campus 802.11 infrastructure. They compared two different campus networks using similar analysis methods and find log normality is prevalent in the aggregate traffic load of APs in both campuses. Moreover, they observe a correlation between the number of associations and traffic load at APs. Authors in [105] presented a session-level and follow-level modeling of traffic in a campus 802.11 network. They modeled the session arrival process at AP as a time-varying Poisson process. They observe the arrival of a user session at AP triggers the arrival of a group of flows that form a cluster process, which can be described by a cluster Poison process. Also, they observe session arrival times at

APs are uncorrelated and exponentially distributed with a mean equal to unity; while the Bi-Pareto distribution yielded the best fit for the number of flows per session, a lognormal model provided the best fit for flow inter-arrivals within sessions. Authors in [106] modeled traffic workload in terms of wireless sessions and network flows, using buildings as basic entities for traffic demand. They find that traffic and roaming patterns varies across various times (hour, day, week) and spatially correlated in terms of building and building-type. Moreover, sessions for the traffic non-stationarity in time followed a time-varying Poisson process, and also, the in-session number of flows and the flow sizes can be well approximated by the Bi-Pareto distribution; whereas the Lognormal distribution was the best fit for in-session flow inter-arrivals out of a set of common distributions including Weibull, Gamma, and Pareto. In [107], authors characterized static and roaming traffic flows at APs from collected 802.11 usage data. They look at TCP flows at each AP and observed their inter-arrival time and duration, and find that Weibull regression model is able to accurately approximate the arrivals of flows at individual APs; on the other hand, both roaming flows with less than five handoffs and the number of handoffs followed Geometric distribution.

Along the same direction of research, authors in [108] analyzed traffic characteristics of a high-speed wireless Internet access sessions. By observing session lengths and traffic volumes, they find that the longer the session the higher the up- and downloaded traffic volumes, the downloaded traffic tend to increase with the increase in uploaded traffic, and in overall, these underlying relationships can be captured quantitatively by the trends determined by the power-law and linear regression. Authors in [109] assessed several network usage metrics related to AP utilization including traffic load. They observe the residential and academic sectors are responsible for most of inbound and outbound traffic load, with nearly 80% of the total. Also, sessions of long duration depends on the AP location (e.g. home or research center), and sessions of short duration occurs mostly when many users are associated to APs.

Similar to the models for association durations and registration patterns at AP in the previous subsections, these proposed traffic models can be influenced by the traffic resulting from the Ping-Pong connections. Since, most of these models consider total traffic generated at AP throughout users-AP association regardless of the type of connections.

3.2.7 Public 802.11 Infrastructure Usage

There exist various efforts in the literature where researchers attempted to analyze usage patterns of the publicly deployed 802.11 infrastructure. Their main goal in these works is to highlight global properties of the infrastructure usage, similar to those explained throughout subsection 3.2.1 to 3.2.6. For example, authors in [110, 111] studied the usage of the Google WiFi network deployed in Mountain View California based on client device type i.e. traditional laptop users, fixed-location access devices, and PDA-like smartphone device. They analyzed usage activities in terms of traffic demands, and mobility of only active clients, as they roam through the city (in their settings client is considered to be active if it sends at least one packet per second during a 15-minute reporting interval, hence trace records have a granularity of 15 minutes). Results shows Google WiFi network has a substantial daily user population, with weekend usage lower than weekdays use; 35% of all devices associate with only one AP, whereas smartphones frequently associate with a large number of APs, and the traffic varies substantially among the different populations with the predominance of HTTP, peer-to-peer, and other TCP traffic. In another effort, authors in [112] examined five weeks traces from the Verizon Wi-Fi Hotspot in Manhattan. In their analysis, they find most clients used the network infrequently and visited few APs, and usage of the network display a strong diurnal usage pattern with weekly trend. The hotspot APs were not that busy even during peak usage periods, and on average the aggregate usage is higher on week-days than on weekends. Authors in [113] examined a 5 month long traces collected by a wireless network service provider operating hotspots in restaurants, serviced apartments, hotels, and airports all over Australia. They analyzed number of sessions established, session durations, and traffic for different user accounts on hourly, daily, and weekly basis. They observe highest user activity during the night with busy hours occurring between 8 p.m. and midnight, whereas the daily activity is roughly uniform for all days of the week. The average session length is about an hour and most of the timed out sessions were idle when implicitly disconnected, on the other hand, the average hourly traffic is high, particularly when most sessions are active, while daily traffic is asymmetric with outbound traffic less than inbound traffic.

In the same line of research, authors in [114] presented a quantitative analysis of the usage of a large public wireless local area network that provide both indoor coverage at a university campus and selected public city premises, as well as outdoor coverage in a city center of Oulu in Finland. They find that university usage is more on office hours and weekdays, while the city usage is higher during the evening hours and weekends; additionally, city usage is higher than university usage in terms of the aggregate traffic volume. Analysis of user mobility indicated users are not very mobile, as less than 10% of the sessions involved spatial movement of at least 50 meters, while 60% of the stations using the network appear to have a home location (access point), where they spent about 80% of their total time. In a slightly different attempt, authors in [115] presented results of Internet traffic measurements of a commercial broadband wireless access network for home users. Analysis of data collected from 250 households reveal that P2P file sharing traffic is used almost all day long, whereas the amount of web and streaming traffic increases in the evening hours with a peak at 19:00 o'clock. A further analysis of the P2P traffic shows that BitTorrent is about 56% and eDonkey is about 41 %; furthermore streaming traffic consumed about 22% of the total traffic besides web and P2P. In [116], authors highlighted the challenges of establishing a global hotspot infrastructure both technical and deployment-related challenges. This is in order to define a viable hotspot business model that would be able to cater and provide added value for all its stakeholders: the end user, the network service provider, and the building and premise owners. They proposed several issues that need to be considered for it to be a reality, for instance issues related to authentication, security, coverage, management, location services, billing, and interoperability.

The usage characterizations of the above mentioned public 802.11 networks produced almost similar results. On average aggregate usage of the network displayed diurnal usage pattern with weekly trend, and higher network usage was noted on weekends than on week-days, which is contrary to a campus setting [56 - 64]. Most users have a home location (AP) with some visited few APs, and traffic is dominated mostly by HTTP and peer-to-peer applications. However, for most of these works the focus is on providing statistics about infrastructure usage, and mobility models of individual and group of users rather than probabilistic models for overall access point usage characterization e.g. probabilistic models for predicting future usage of an AP based on the current and previous usage patterns of the AP.

3.3 Anomaly Detection

3.3.1 Overload Detection

The IEEE 802.11 MAC protocol originally was designed to offer asynchronous best-effort service to stations within 802.11 network systems. In the sense that each station associated to 802.11 AP is granted the same transmission opportunities similar to other stations in the long term. This means if the number of associated stations to an AP increases the throughput per station is likely to decrease drastically. Increase in the number of associated stations to an AP can results into queues buildup in both the stations and the access point, leading to loss of frames, which could be detrimental to the overall 802.11 performance.

In connection to this, there exist various efforts in the literature that tried to investigate the underlying properties of overloaded 802.11 networks from collected 802.11 usage data. For example, authors in [117, 118] analyzed variations in the link-layer properties during network overload period. In [117], authors observe that the overloaded wireless network is characterized by extensive medium occupancy, high traffic, frequent bit errors, numerous retransmissions, and significant data rate variations. In [118], authors find that the use of RTS/CTS denies nodes from gaining fair access to a heavily congested channel, and the use of rate adaptation in response to network congestion generated a lot of handoffs which impact negatively 802.11 performances. Authors in [119] shows that, under overload conditions stations only maintain a short association period with an AP, and repeated association and reassociation attempts due to lost connections are common phenomena even in the absence of mobility. Their analysis of handoffs shows that stations’ throughput suffers drastically following each handoff, leading to suboptimal network performance. In [120], authors find that the increase in packet losses necessitate stations to initiate a handoff in search of a better AP in their vicinity, and the use of channel busy time allows in differentiating packet losses due to overload and those due to poor link quality.

These works do not propose any mechanism or method for overload detection or prevention. They only analyzed a section of 802.11 network usage data where the highest number of users was observed in the 802.11 APs, and ignored the fact it is also possible for some few users to also cause AP overload.

To minimize user connectivity disruptions (repeated handoffs, bit errors, retransmission) during overload situations and for proper balance of network usage, several techniques have been proposed in the literature. For example, authors in [121, 122] presented a distributed admission control that limits the possibility of AP overload due to a high arrival rate of network flows. In their admission systems, each station perform a short probe that estimates MAC service time, the offered load, and transmission rates in an AP to determine if AP has enough capacity for a new connection. Thus, the flow is admitted only if the estimate is below a predefined threshold, otherwise the proposed admission controls block stations from initiating new sessions. In [123], authors proposed a queue-based user association management system to mitigate the impact posed by arrival of flash crowds and presence of high concentrations of users in 802.11 networks. Their propose system maintains a queue of users that request network access, and grants 802.11 network accesses only to a limited number of users at a time. One limitation of these schemes is that they require modification at APs (e.g. scheduling and bandwidth partitioning mechanisms) which cannot be easily adopted in large-scale 802.11 networks.

Along the same direction of research, authors in [124] proposed a load-balancing scheme in which agents' runs on each access point and periodically exchange load information (AP throughput) to determine whether AP is overloaded, balanced, or under-loaded. In their approach, access points which are overloaded forces handoff, and only lightly loaded access points must accept roaming stations. In [125], authors proposed a dense access points deployment architecture (DenseAP), where a central controller collects throughput information from all APs before deciding which AP should each client associate with. In addition, the central controller also decides on the assignment of channels to APs and also performs periodic load balancing to reallocate clients from APs with significant load to nearby APs with light loads. Authors in [126] proposed a system for automatic access point discovery and selection called Virgil. In their system station first scans all available APs in a given location and quickly associates to each AP, and runs a series of tests to estimate the quality in terms of throughput and delay of each AP's connection to the Internet before choosing best AP to associate with.

However, these proposed schemes improve aggregate throughput, fairness, and delay at 802.11 AP, but they require certain support and cooperation between clients' and access points which cannot be easily applied in large-scale 802.11 networks, given the higher number of users, access points, and network locations. Furthermore, the focus in most of these works is in controlling the number of newly arrived users that requires admission to 802.11 AP, and gives less importance to the detection of AP overload situations resulting from users who are already associated to the 802.11 AP.

In a slight different effort, authors in [127] presented a load balancing technique similar to cell breathing in cellular networks that controls AP's coverage range (cell size), by dynamically changing the transmission power of the beaconing messages of AP. They develop numerous polynomial time algorithms to help find the optimal beacon power settings that could minimize the load of the most congested AP, in terms of number of users and traffic demands. In [128], authors modelled access point selection as a game and assess the impact of user dynamics on arrival/departure patterns system behavior. They find that their proposed access point selection algorithm bring the game to a Nash equilibrium in a matter of single iteration, as each user makes a single selfish AP selection based on the current state of the system.

One limitation of these proposed schemes is that users are likely to be shifted between APs frequently, owing to the time varying nature of the wireless medium and bursty nature of wireless users' traffic. These two inherent wireless properties are likely to influence changes in traffic load and delay in the AP at any moment in time, and as a consequence of that users will be shifted to other APs each time any of these happens.

3.3.2 AP Halt/Crash Detection

Due to time varying nature of the wireless medium and usage pattern behavior of a large-scale 802.11 infrastructure, it is possible one or more access points at any given moment in time to face problems and stop running (i.e. halt or crash). As a consequence of AP halt/crash, clients' connections and throughput will be severely impaired, leading to significant amount of intermittent sessions as user clients will keep probing and searching for good connections in their vicinity, repeatedly. This can cause wireless medium to be busy, due to extensive medium occupancy by management frames, and therefore preventing other stations from accessing the medium. To detect such failed AP by probing the wireless interface of all APs in the infrastructure proves to be a labor-intensive and inefficient task. The ability to automatic detect AP halt/crash is valuable for effective management of large-scale 802.11 infrastructures.

In connection to this, there exist few efforts that tried to detect halted/crashed AP from 802.11 measurement data. For example, authors in [129] proposed an algorithm that exploits device mobility to detect presence of faulty APs in an 802.11 network. The main assumption in their algorithm is that the longer the time an AP does not register events, the greater the probability that particular AP is faulty (i.e. halted/crashed). In our work we consider the absence of registered events in a specified time after AP abrupt ending of connections and not just after last registered association event at an AP. In [130], authors proposed architecture to improve fault tolerance during access point failures, where centralized management system (MS) collects measurement from all APs on a timely basis. In their approach whenever a miss reading (RSSI) measurement from AP is detected, a centralized Management Station (MS) polls that AP to confirm such AP failure, and remotely sets the new configuration in other working APs. While this proposed scheme is able to detect out of service AP (i.e. AP which is not responding to status probes because of the problem in the wired link), but it fails to detect AP that broadcast wireless beacons and yet cannot allow devices to associate.

Authors in [131] proposed and evaluated a technique called Client Conduit, which enables bootstrapping and fault diagnosis of disconnected clients around APs. Their solution focuses primarily on the use of controller with assistance of a diagnosis server to detect and self-diagnose disconnected clients around halted or crashed APs. In their approach, clients are augmented to start an ad-hoc network whenever an AP is faced with a problem. Authors in [132] proposed a distributed self-diagnosis protocol where nodes test each other in the network. In their approach, each mobile node sends a task to its immediate neighbor nodes to determine if they are running or failed, then the output of this test is shared to all other nodes for attaining a global view of the network status.

The focus in these works is in supporting disconnected clients to not associate with a halt/crash AP, but for that to be possible clients require cooperation among themselves and support from a third party device such as diagnostic server or management station. Instrumenting devices (APs and Clients) and introduction of third party devices in large-scale 802.11 networks can be difficult and expensive owing to higher number of users and APs.

3.3.3 Interference Detection

Interference in 802.11 networks in most cases is caused by the broadcast nature of wireless links where transmission in one link of the network is able to interfere with the transmissions in other neighboring links. In addition to time varying nature of the wireless medium, interference in 802.11 networks can also be caused by other radio waves in the same frequency range. During interference data hardly make it through the air, in most cases requiring lots of retransmissions which results in overall 802.11 performance degradation. There exists a large body of work about interference in the literature, where researchers studied the effects of interference from different perspectives before suggesting ways and techniques to detect and mitigate its potential impact.

For example, authors in [133] studied the impact of interference in chaotic 802.11 deployments on end-client performance. Result shows the performance of end-clients throughput suffers significantly in chaotic deployments, as most APs are not configured to minimize interference with their neighbors. In [134], author propose methods that include intelligent frequency allocation to APs, load-balancing of user associations to APs, and the use of adaptive power control to APs to mitigate the effect of interference in dense 802.11 deployments. Authors in [135] examined QoS characteristics for mobile-WLAN devices, in terms of the measured throughput, when multiple m-WLANs use different channels at different geometric distances. They find the effect of the channel distance on the throughput is non-uniform in nature, with some distances resulting into more than the expected channel interference. Authors in [136] studied the impact of RF interference on 802.11 networks, from range of devices such as Zigbee and cordless phones that crowd the 2.4GHz ISM band, and also from devices such as wireless camera jammers and non-compliant 802.11 devices that disrupt 802.11 operations. They experimentally confirm that changing 802.11 operational parameters such as clear channel assessment (CCA) threshold and rates is not effective at withstanding interference compared to moving to a different channel.

Along the same direction of research, authors in [137] proposed an online interference estimation mechanism (PIE), implemented at the AP with a central controller placed at the wired network to observe ongoing traffic at different APs. They demonstrate that PIE is able to diagnose interference as well as certain performance issues such as hidden terminals and rate anomaly in real deployments. In [138, 139], authors proposed tools that provides a time-domain view of how the medium is used in a given 802.11 channel. In [138], authors demonstrate that physical layer properties such as bit error patterns and medium busy times can be exploited to identify the cause of interference from non-802.11 devices. Whereas in [139] authors demonstrate that when the airspace is congested, changes in a victim node's throughput are more closely related to its interferers than other devices.

Other interesting research include [140], where authors investigated the problem of learning the graph structure based on network traffic transmission patterns, especially information about successes and failures in transmissions. They find that the networks with sparse interference patterns can be quickly identified by their approach than those with dense interference patterns. In [141], authors presented a general model for error probability and throughput of packet transmissions on a small wireless networks. In their approach, senders and receivers communicate at short distances and potential interferers are slightly far away.

In the same line of research, authors in [142] proposed a model based on a Markov chain to capture the interaction between different senders and receivers in heterogeneous multihop wireless networks. The proposed model takes input traffic demand and received signal strength (RSSI) between pairs of nodes to estimate interference among an arbitrary number of senders. Authors in [143, 144] presented an approach based on Hidden Markov Model to estimate the interference between nodes and links by passive monitoring of wireless traffic. They first identify the interference relations between nodes and links, and model the 802.11 MAC as a Hidden Markov Model (HMM) to infer pairwise interference. The proposed technique in [143] was able to estimate non-binary pairwise interference, but failed to infer aggregated interference from a set of nodes. Whereas the proposed technique in [144] was able to infer the carrier-sense relationship between network nodes, but failed to detect selfish carrier-sense behavior of network nodes.

Most of these approaches considered just one or few set of APs in their evaluation, or make use of dedicated sniffers and controllers to help capture network traffic for potential detection of interference patterns at the sender, link, and receiver. This again is not scalable due to a large number of users, APs, and network locations that exists in these 802.11 large-scale deployments.

3.3.4 Rogue AP Detection

A rogue access point (Rogue AP) is an 802.11 access point that has been installed on a wired side of an 802.11 network, possibly without explicit authorization from the network administrators. This can be accomplished in either of two ways 1) it could be naively installed by a legitimate user who is not aware of its potential security implications, or 2) it could be intentionally installed as an insider attack. A rogue AP can pose significant security threat to a wired part of 802.11 networks; since it is possible through it to provide a backdoor into an 802.11 network to outsiders (i.e. illegitimate users). Moreover, rogue AP can potentially affect network connectivity by interfering with nearby legitimate APs, leading to user intermittent connectivity and in some cases authentication failure at APs.

The first step in the controlling of a rogue AP and its potential impact is to detect its existence. There are numerous efforts in the literature that try to automate rogue AP detection beyond the monitoring components normally used by network administrators to listen to wireless frames from all access points and to compare to the prerecorded list of legitimate APs. For example, authors in [145] proposed a centralized system that collects data transparently at data aggregation point such as a router or a gateway to detect the possibility of existence of rogue APs. Their approach first discovers all wireless stations associated in the network and verifies their authorization; thus any unauthorized wireless station found in a given AP indicates that the attached access point must be a rogue AP. In the same line of research, authors in [146, 147] used a packet analysis method to detect rogue APs. Their proposed method compares the gateways and the routes that each packet travels to determine whether an access point is legitimate or rogue. In [148], authors proposed a framework that separates frames into IP and TCP components; allowing for information such as client MAC addresses, SSID, channel assignment, encryption status, and beacon interval to be analyzed for potential events generated by a rogue AP.

In another effort, authors in [149, 150] examined the round trip time (RTT) variation of network traffic resulted from the DCF, link-layer retransmissions, contention, coupled with a list of authorized APs and their respective switch ports to detect presence of rogue access points on a wired link. In [151, 152, 153], authors proposed a timing-based client-centric approach that uses the round trip time between the user and the DNS server to detect whether an AP is a rogue AP or not. In their scheme, the user will send a series of DNS requests and measure the RTT from the local DNS server; an abnormal RTT will be considered coming from rogue AP and therefore detected by the user.

These works do not attempt to detect connectivity anomalies resulting from the existence of a rogue AP. Detection of rogue AP can minimize cases of user intermittent connectivity and user authentication failure in the network. In addition to time varying nature of the wireless medium and RF holes, the aforementioned user-related anomalies have potential to manifest themselves possibly due to interference effects caused by rogue APs in these large-scale 802.11 networks. We do not detect Rogue AP in this thesis, but we focus on detecting the user-related anomalous patterns of user’s intermittent connectivity and user’s authentication failure at APs that may perhaps be caused by Rogue AP.

3.3.5 Performance Anomaly Detection

In addition to RF effects such as exposed and hidden terminals, the capture effect, and fading; there also other performance anomalies which contribute to overall connectivity degradation of 802.11 networks. In this section, we try to enumerate few important works that analyzed performance anomaly from 802.11 measurement data.

This is for the reason that usually connectivity anomaly in these large-scale 802.11 networks manifest as a result of performance anomaly.

For example, authors in [154] analyzed measurements data collected during the SIGCOMM 2004 conference, aiming to understand performance of 802.11 in real deployments. In their analysis, they find that the overhead of 802.11 is high, with only 40% of the overall transmission time being spent in sending original data, while most of the remaining time is consumed by retransmissions due to packet losses caused by varying signal strength. In [155], authors showed under the conditions of high medium utilization and packet loss, handoffs are incorrectly initiated in 802.11 networks. Also, they observe a significant fraction of these handoffs were unnecessary and to the same AP.

Along the same line of research, authors in [156] proposed a model of MAC protocol behavior to infer delays such as AP queuing, backoffs, and contention from 802.11 measurements. They find that no one anomaly, failure, or interaction is singularly responsible for these issues. In [157], authors showed that in an unplanned multi-cell network the collision rate increases significantly, the number of backlogged stations equals twice the number of active access points, and clients such as VoIP users largely experience substantial performance degradation. In [158], authors showed that increasing the offered load to the access point's Ethernet interface does not always increase the downlink throughput; however, few access points present a downlink throughput reduction when the offered load exceeds their bridging capabilities.

Focusing on AP placement, authors in [159] studied the impact of access point (AP) configuration and placement on the aggregate throughput of 802.11, without using any mathematical model or radio channel characterization using only a pair of wireless laptops and APs. Their result shows that appropriate AP configuration and placement can have significant impact on the overall 802.11 network performance. In [160], authors proposed a mechanism that resolves throughput imbalance among Basic Service Sets (BSSs). In their proposed approach, a central controller arbitrates the wireless channel occupation of APs in the large-scale 802.11 networks. Authors in [161] presented performance analysis of AP connections from highly mobile clients, through various indoor/outdoor experiments and analytical model. Experimental results show that using multiple APs on a single channel achieves higher throughput than scheduling on multiple channels. In [162], authors provide an approach for optimizing AP deployment with very little signal to interference noise ratio (SINR) degradation. Their proposed approach examined the saturated throughput, user QoS, and energy consumption performances of 802.11n networks under error-prone channels.

Focusing on RF effects, authors in [163] analyzed the capture effect both as a function of delay and signal strength and showed that it is quite strong, especially at lower transmit rates. They also find that off-channel interference reception behavior is rather poor and that hidden nodes are uncommon in dense wireless networks. Authors in [164] showed that, in a network with more than two nodes, the capture effect plays a significant role, and affects both the fairness and failure rates in the network. In another effort, authors in [165] explore 802.11's behavior under the capture effect through examination of the interaction between the PHY and MAC operation. In their testbed experiment they show that the capture effect can in fact be exploited to improve the overall throughput of the network. Authors in [166] proposed an analytical model to evaluate the effect of hidden station on both non-saturation and saturation performance of the Distributed Coordination Function (DCF). They find that hidden stations usually cause collisions which results in significant degradation of the network performance.

There exist several monitoring systems that tried to diagnose performance anomaly at different layers of 802.11 protocol stack. For example, authors in [167] proposed MOJO, a system that detect hidden terminals in the network and activity from terminals experiencing capture effect at the granularity of PHY layer. Also, MOJO include algorithms that detect noise due to non-802.11 devices and long term anomalous signal strength variations at the AP. Authors in [168] presented a system called Jigsaw that uses multiple monitors (sniffers) to provide a single unified view of physical, link, network, and transport layer activity on an 802.11 network. They used their proposed system for a cross-layer detection of performance problems such as lost frames and packets due to weak signal strength and collisions at APs. Authors in [169] proposed a solution WiFox, which adaptively prioritizes AP’s channel access over competing stations to avoid traffic asymmetry. They show that their proposed solution can alleviate the problem of performance loss due to rate-diversity/fairness and degradation due to TCP behavior. Authors in [170] proposed DOMINO, a system for detection of greedy behavior in the MAC layer. The proposed approach consists of six tests, where test 1 is designed to examine scrambled CTS/ACK/DATA frames, and the others (tests 2 through 6) are designed to detect compromised protocol parameters. Authors in [171] presented a diagnostic system that employs trace-driven simulations to detect faults in multihop wireless networks. Using their approach, they were able to diagnose performance problems caused by packet dropping, link congestion, external noise, and MAC misbehavior. Authors in [172] built MODI-embedded wireless APs that cooperate to detect and troubleshoot performance problems such as connection failure, throughput degradation, and transmission delay. From testbed experiment, authors show that MODI can achieve reasonable accuracy when diagnosing both individual and simultaneous faults.

On a slight different direction of research, authors in [173] proposed MAP, a system for detecting deauthentication/disassociation management frames initiated by attackers to terminate 802.11 authentication/association procedures. Evaluation results from a testbed indicate MAP was able to detect with reasonable accuracy spoofed authentication/association procedures geared to disrupt wireless connectivity between the clients and target AP. In [174], authors identified several anomalies in MAC protocol such as excessive retransmissions of some management frames (e.g. Probe Response (64%), Reassociation Request (25%), and Power-Save Poll (13%)) which could be detrimental to network performance. In [175], authors used K-mean clustering algorithm to separate time intervals of both normal and anomalous traffic, and use cluster centroids to detect anomalies based on distance calculations.

While these proposed frameworks provide different monitoring solutions for different 802.11 performance anomalies, their main limitation is that the deployment scenarios investigated in these works are too simple to reflect the real characteristics of large-scale 802.11 deployments. The performance of each proposed method is evaluated based on either a small scale network, on a testbed, or through simulations. Therefore, little is known about how these proposed schemes can be able to scale-up to large 802.11 networks with minimum overhead. Also, a guarantee of their actual performance when tested with realistic workload from real large-scale 802.11 deployments is not known.

3.3.6 Network Management Tools

Nonetheless, there exists several open source as well as commercial products that aid 802.11 network management tasks. For example, Network and Systems Management (NSM) [176], Wireless Security Auditor [177], AirMagnet [178], AirDefense [179], and those that based on sight survey mentioned in [148][131][192].

The main limitation of these tools is that their main focus is to provide statistics and graphs of aggregate 802.11 AP usage, rather than detecting anomalies and to establish their respective true nature. While statistics such as number of retransmissions, number of dropped frames/packets, TCP loss ratio, and packet throughput at the AP, helps network administrators to understand general usage state of the 802.11 infrastructure, but depicting anomalies and to establish their possible root causes from these low-level statistics remain to be a challenging task. Some products like Wireless Security Auditor [177] monitor a set of measured indicators to determine threshold violations (alarms); although, a network manager must then analyze alarms in order to establish possible causes of potential malicious attack.

Other network management products like AirMagnet and AirDefense ([178, 179]) make use of deployed hardware sensors to collect packets across different parts of the 802.11 network before sending to a central dedicated server for fault analysis. However, to deploy hardware sensors to capture a complete view of the large-scale infrastructure network can be increasingly difficult and expensive.

3.4 Analysis

In this section, we provide discussion of the related work as far as the problem of usage modeling and anomaly detection in large-scale 802.11 networks is concerned. Our discussion is based on table 1.Table 1 depicts different types of works analyzed in the related work pertaining to the problem of usage modeling and anomaly detection in 802.11 networks. The columns in table 1 represent key features employed by these works. We will analyze features provided by the related work in comparison to the features pertinent to our proposed framework later in this section. The following is the description of each of the feature in table 1.

- Models for AP Usage Characterization: This feature aims to find out if the type of work analyzed has tried to propose models that depict access point usage characterization, and not just statistics about AP usage. For example, if the work has provided models for characterizing users-AP association events, user mobility, user dwell time at AP, and traffic at AP.
- Based on Total Association Events at AP: This feature aims to understand that if the proposed model is based on aggregate associations from all users associated to AP, and not just from certain individuals or group of users' association events at AP. For example, models based on hourly or daily total number of users associated to an AP and their respecting access duration and traffic.
- Effect of Ping-Pong Sessions in Modeling: This feature aims to understand that if the work analyzed has considered the Ping-Pong sessions in developing AP usage models, and not just ignored their effects in constructing usage models. For example, if the work has considered removing repetitive connections from a single user at various distant APs in a short time interval (few seconds or minutes), and their respective traffic.
- Probabilistic Nature of the AP Usage Models: This aims to find out if the proposed AP usage models in the work are probabilistic in nature. In the sense that, if the proposed models in the work are generative and that there is a possibility of generating synthetic traces to use, for example in network simulation tools for evaluating applications and protocols in large-scale 802.11 infrastructures.
- Suitability of the Models for Capacity Planning: This feature aims to understand that if the proposed models in the work can be used to understand and predict overall access point usage. For example, if the work has considered models that can predict aggregate hourly or daily users-AP association events, access duration, and traffic at AP regardless of unique nature of wireless user associations, their similarity, and encounter patterns at AP.
- Ease of data Access (low cost data): This feature aims to understand that if the data used in the evaluation of the work can be easily obtainable. For example, if the data set used in the work can be easily accessible in all parts of the network at low cost. For example, if the type of data is able to capture all users-AP

association events from all APs, and has ability to span seconds, minutes, hours, days, weeks, months, or years.

- Tested on Large-Scale Network: This feature aims to understand that if the data used in the evaluation of the work is actually from the large-scale network, and not synthetically generated by a simulator, or from a testbed, or from a small network. For example, if the data used is not from a network with just few APs and users.
- Ease of Approach for Anomaly Detection: This feature aims to find out if the proposed approach for anomaly detection in the work does not involve instrumenting devices (e.g. APs and clients) or introduction of a third party device in the architecture (e.g. diagnostic servers, sensors, controllers, and management stations). Also, if the proposed approach is able to scale-up to a large-scale network with minimum cost and overhead. For example detection of anomalies solely based on the analysis of user association patterns at APs.
- Models for Estimating Anomalies in AP Usage: This feature aims to understand if the proposed approach for anomaly detection includes statistical models for charactering occurrences of anomalies in the usage of APs. For example, statistical models that estimate occurrences of anomalies such as AP interference, AP overload, and AP crash in the usage of APs, with respect to aggregate network usage in terms of the total number of sessions. For example, linear regression models or continuous probability distribution models for estimating anomalies in relation to the aggregate 802.11 network usage, in terms of the total number of sessions in the network.

			Features								
S/No	Type of Work	REF#	Models for AP usage Characterization	Based on Total Association Events at AP	Effect of Ping-Pong Sessions in Modeling	Probabilistic Nature of the AP Usage Models	Suitability of the Models for Capacity Planning	Ease of data Access (low cost data)	Tested on Large-Scale Network	Ease of Approach for Anomaly Detection	Models for Estimating Anomalies in AP Usage
1	**General Statistics about 802.11 Usage**	51-55	x	x	x	x	x	√	√	x	x
2	**User Mobility Modeling**	56-64	√	x	x	x	x	√	√	x	x
3	**Location Prediction**	65-70	√	x	x	x	x	√	√	x	x
4	**Ping-pong effect and mobility**	71-72	√	x	√	x	x	√	√	x	x
5	**Human Physical Movement**	73-80	√	x	x	x	x	x	x	x	x
7	**User Encounter Patterns**	81-88	√	x	x	x	x	√	√	x	x
6	**User Registration Patterns**	89-95	√	√	x	x	√	√	√	x	x
8	**User Access Durations**	96-100	√	√	x	x	√	√	√	x	x
9	**Traffic Characterization**	101-109	√	√	x	x	√	√	√	x	x
10	**Public 802.11 Infrastructure Usage**	110-116	√	√	x	x	x	√	√	x	x
11	**Overload Detection**	117-120	x	√	x	x	x	√	x	x	x
		121-128	x	x	x	x	x	x	x	x	x
12	**AP Halt/Crash Detection**	129	x	√	x	x	x	√	x	x	x
		130-132	x	x	x	x	x	x	x	x	x
13	**Interference Detection**	133-135	x	x	x	x	x	√	x	x	x
		136-144	x	x	x	x	x	x	x	x	x
14	**Rogue AP Detection**	145-147	x	√	x	x	x	√	x	x	x
		148-153	x	x	x	x	x	x	x	x	x
15	**Performance Anomaly Detection**	154-157	x	√	x	x	x	√	x	x	x
		158-175	x	x	x	x	x	x	x	x	x
16	**Network Management Tools**	176-177	x	x	x	x	x	√	√	x	x
		178-179	x	x	x	x	x	x	x	x	x

Key	
Feature Provided by the Work	√
Feature not Provided by the Work	x

Table 1. Features of the related work

Discussion: From table 1, it’s clear that the majority of works in literature about 802.11 usage focused on modeling user mobility, user encounter patterns, user registration patterns, user access duration, and traffic at APs (item 1-10). For most of these works, especially item 2-6, the focus is to obtain particular statistics about individual users or group of users associations at AP, and to establish a model based on these quantities. This is particularly important if the goal is to understand user behavior, but if the aim is to understand access point usage behavior, then models that consider total user association events at AP regardless of movement patterns of individual users or group of users or their similarity in terms of associations and encounters patterns at AP can be increasingly beneficial to network administrators as far as capacity planning and resource provisioning of a large-scale network is concerned. For example, works about user registration patterns, user access durations, and traffic (item 7-9). Limitation of these proposed models is that they do not exclude association events and traffic resulting from the Ping-Pong sessions, as it was evident in the related work that removing Ping-Pong sessions in constructing statistical models can improve their fidelity significantly [71, 72]. This means models that get rid of the Ping-Pong sessions and at the same time capable of characterizing overall access point usage can be of great benefit to 802.11 administrators, than just those that consider total user associations, access durations, and throughput regardless of the type of connections established at APs. These observations are also applicable to the works about public 802.11 infrastructure usage (item 10). The only exception is on the works about general statistics (item 1) where no attempt was made to model 802.11 usage.

On the other hand, most proposed approaches in the literature for detecting anomalies in 802.11 networks such as AP overload, AP crash, and AP interference (item 11-13) requires modification at either client or AP [121-128], or implementing cooperation between clients [130-132], and introduction of a third party device such controller, sniffers, and sensors [136-144]. However, instrumenting devices (clients and APs) and deployment of a third party device such as sniffers, controller, and sensors in 802.11 large–scale infrastructures can be increasingly difficult and expensive. Other approaches for anomaly detection, for example [148-153] and [158-175], were evaluated based on a small scale network, on a testbed, and through simulations, hence there is no guarantee on how these frameworks are able to scale-up to large networks with minimum cost and overhead. This is in addition to uncertainty about their actual performance when tested with real workload from a large-scale 802.11 deployment. There works for anomaly detection in 802.11 networks based on low cost data such as RADIUS session data and SNMP data, for example [117-120], [129], [133-135], [145-147], and [154-157]. However, most of these works examine only few set of specific APs instead of all APs in the infrastructure. Hence their scalability is also not guaranteed.

We conclude that there is no solution in the related work that supports all features depicted by table 1. As it was evident in table 1, most works in the literature supports only a subset of these features. Our proposed framework in this thesis is the only solution that supports all features provided by table 1. Moreover, there also features peculiar to our proposed framework which are not supported by any of the previous work. These are columns in table 1 filled completely with marker “x”. Specifically, none of the work in the literature has considered probabilistic generative models for automatic access point usage characterization based on the aggregate users-AP association events from all users on all APs. In addition, none of the work has considered analysis of session endings at 802.11 AP for ease detection of anomalies in the usage of APs e.g. AP halt/crash, AP overload, AP interference, interference across the vicinity of an AP, AP persistent interference and, user authentication failure and user intermittent connectivity to APs. Furthermore, none of the work has considered statistical models for estimating occurrences of the aforementioned anomalies in AP usage with respect to aggregate network usage in terms of the total number of sessions.

Unlike all related works, our proposed framework in this thesis is the only solution that performs both performance and fault management using easily accessible low cost data collected from the real large-scale 802.11 infrastructure. The ease of our proposed framework makes it a viable option for immediate adoption, complementing the existing network management approaches in large-scale 802.11 networks

3.5 Conclusion

In this chapter we analyzed relevant works pertaining to the problem of usage modeling and anomaly detection in 802.11 networks. We presented more insights on the techniques and features employed by most existing works in the literature. Despite all these works in the literature, to the best of our knowledge this is the first attempt to derive generative probabilistic models of AP usage based on the readily accessible RADIUS session data (i.e. keep-alive event counts generated every 15 minutes for refreshing user-AP connections) and to compare them using the log-likelihood and AIC values figures of merit (chapter 4). This is also the first work to identify a usage pattern namely “abrupt ending” of 802.11 AP connections that happens when all or a significant number of user connections end in the same access point within a one second window (chapter 5). The emphasis on AP usage and the use of probabilistic generative models is better suited to performance management of 802.11 networks, especially capacity planning, and on network simulation experiments for evaluating protocols and applications in these large-scale 802.11 infrastructures. The emphasis on detecting abrupt ending of AP connections and their resulting patterns is better suited to fault management of 802.11 networks, as it may help network administrators to quickly and efficiently detect and fix connectivity problems in the usage of the access points (APs) of the large-scale 802.11 infrastructures (chapter 6).

Chapter 4

Modeling 802.11 AP Usage

4.1 Introduction

In this chapter we motivate and introduce our modeling effort of 802.11 access point (AP) usage. Our goal is to propose generative probabilistic models of AP usage, which embodies the perception of the time that 802.11 users stay associated with APs when accessing network resources of a large-scale 802.11 network. We model usage based on AP daily keep-alive event counts. Keep-alive events are message sent by a mobile client every 15 minutes for refreshing the client's association with an AP. Due to their periodic nature, keep-alive statistics can provide us with the estimates of the time users stay associated with an AP during which they may generate traffic with different profiles. We also employ an above-below AP event count average binary indicator that allows thinking about AP usage modeling in binary terms such as when AP usage is high or low. We train and evaluate our models based on a data set collected from FEUP's 802.11 Eduroam hotspot. The models we present in this chapter are generative, in the sense they can be used to generate synthetic daily event counts for a single AP or a collection of APs. The emphasis on AP usage and the use of probabilistic generative models is better suited for: 1) performance management of large-scale 802.11 infrastructures e.g. models for capacity planning suitable for resource provisioning of the large-scale 802.11 networks, and 2) network simulation tools for evaluating applications and protocols in 802.11 infrastructures.

In this chapter we proceed as follows. We first describe our experimental setup, the choice of data sets, and figure of merit used for evaluating our models in section 4.2. In section 4.3, we propose and evaluate different time-independent and time-dependent models for characterizing AP usage and present their evaluation results in section 4.4. In section 4.5, we evaluate models that consider week structure usage samples of APs. In section 4.6, we present performance evaluation of our extended time-dependent models based on data sets other than the one on which they were trained for. In section 4.7 we present concluding remarks.

4.2 Experimental Setup

4.2.1 Variables

In this work, we use RADIUS authentication data. All 802.11 APs on campus are configured to send 1) log event *'START'* whenever the wireless client authenticates or roams into the network, 2) interim log event *'ALIVE'* every 15 minutes during the entire duration of connection for refreshing wireless clients association to APs, and 3) log event *'STOP'* whenever the wireless clients disassociate or dis-authenticate from the network. All these log events are sent from 802.11 APs and are recorded in a central server running the RADIUS protocol (Remote Authentication Dial In User Service) specifically for authentication, authorization, and accounting as stipulated in RFCs 2865 and 2866 of IETF [35, 36].

The variables we want to model in this work are the AP daily keep-alive event counts, i.e. the series of interim updates generated periodically between the start of user connection at an AP until the end of user's connection. Owing to their periodic nature, keep-alive statistics can tell us estimates of the time users stay associated with an AP, during which traffic of different profiles can be generated. We are using daily statistics because of the natural daily behavior cycle of users, especially on campuses. Previous studies on user behavior have indicated that the number of clients using a particular access point exhibits a weekly periodic behavior with a strong daily pattern, particularly during working days [56 - 64].

Rather than modeling traffic directly, we choose to model event counts based on the following two arguments. 1) There is evidence that user/AP association pattern and thus event counts are correlated with access duration and traffic at APs ([104, 108, 113] and figure 5). This means that event counts can also be used to understand baseline usage for APs, and consequently to support resource reservation in large-scale networks. 2) The possibility of plugging in different traffic models onto our event count models to generate different network traffic patterns according to the needs of simulation experiment or protocol to evaluate.

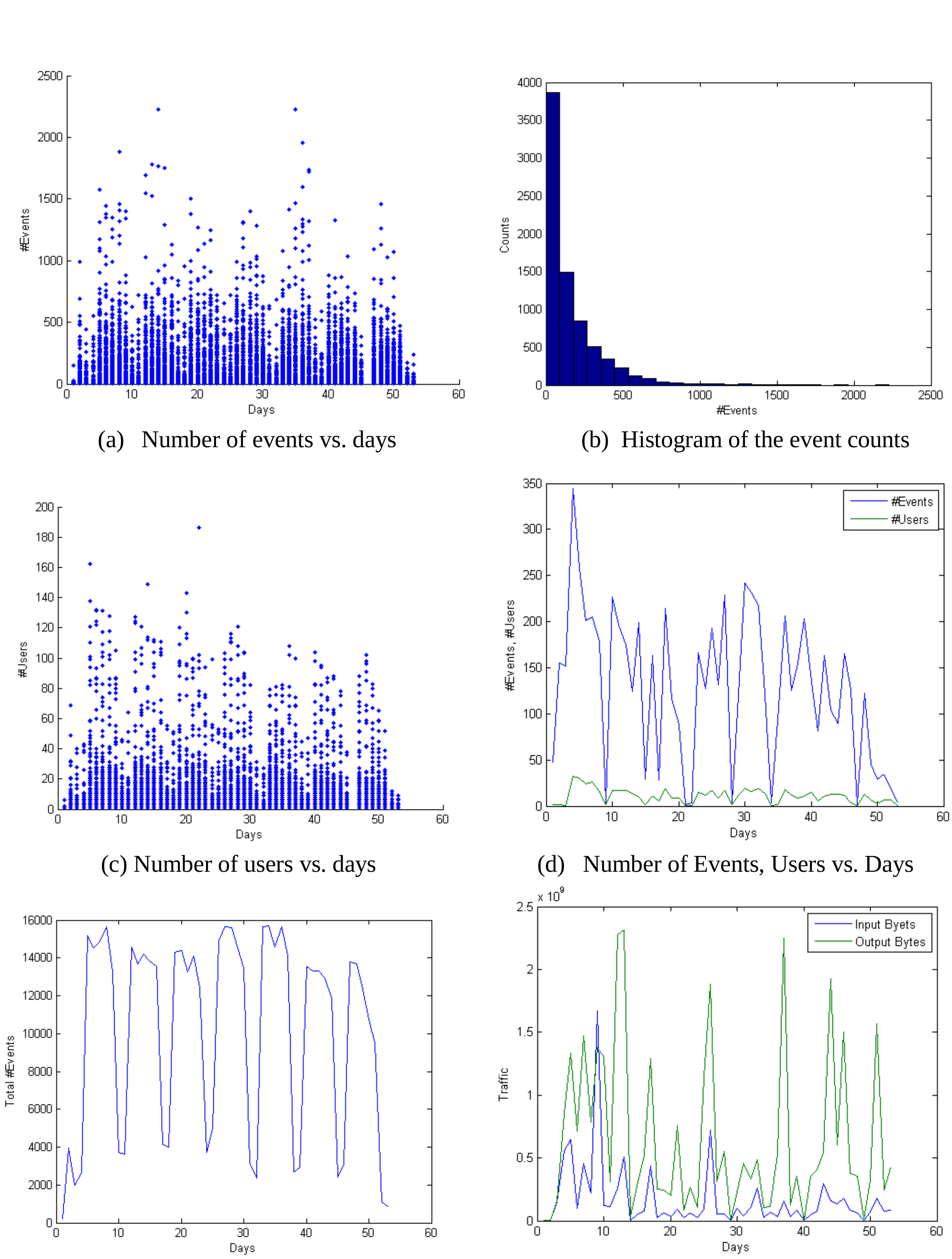
4.2.2 Data Set

Our data set is a 183x53 matrix consisting of daily keep-alive event counts from 183 access points on 53 consecutive days. All our 183 APs are located on campus buildings bearing the same SIID. This allows wireless users to roam seamlessly between campus buildings and campus physical space, since most of these APs also cover outdoor environments. Buildings in our campus are mainly class rooms, theaters, offices, laboratories, canteens, libraries, etc., with the exclusion of residential buildings. The 53 consecutive days chosen for training our models include week-days and week-ends during one semester of the academic year 2009.

Daily event counts in this data set are integers that range from 0 to the maximum of 2229, while the corresponding number of users responsible for these events ranges from 0 to the maximum of 186, (i.e. total number of users observed on campus during the studied trace period). Figure 5 provides more details on this usage information. We observe more usage on the week-days than on weekends. Note the peaks (week-days usage) and the crests (week-ends usage) in plot "*a*" through "*e*" with a periodic (i.e. weekly) repetition. Figure 5(a) shows the number per day from all 183 APs. Figure 5(b) depicts distribution of the daily event counts of APs in the hotspot during the studied 53 days. Figure 5(c) shows the number of users per day from all 183 APs. Figure 5(d) shows the relationship between the number of events and the number of users over the 53 days. Figure 5(e) and figure 5(f) show the total number of events and the total number of traffic for all 183 APs, respectively.

We split the data set and use two subsets for training/testing our models, taking into account that any value over 2/3 is appropriate for training the models [181]. Our training set includes data from the 53 days of 143 randomly selected access points. The test set includes data from the 53 days of the remaining 40 access points that were not used in the training set. We employ AP-based splitting not just because of its simplicity but rather due to the overall goal of our work, which is to understand usage of AP and try to predict daily and weekly usage of APs in the hotspot.

Pearson's Linear Correlation Coefficient: in order to gain deeper insights regarding relationships that may exist between different aspects of 802.11 AP usage in our data set, we employ Pearson's linear correlation coefficient. We check for possible correlation between the number of users and the number of events, and also between the total number of events and the total traffic. We observe positive correlation in each case: 1) for the number of users and the number of events the correlation coefficient is 0.688, which can be visually seen in figure 5(d); while 2) for the total number of events and the total input bytes correlation coefficient result is 0.7063; and 3) for the total number of events and the total output bytes correlation coefficient result is 0.754, portrayed visually in figure 5(e) and 5(f).



(a) Number of events vs. days (b) Histogram of the event counts

(c) Number of users vs. days (d) Number of Events, Users vs. Days

(e) Total number of Events vs. Days (f) Total traffic in Bytes (Input, Output) vs. days

Figure 5. Underlying features of the access points usage pattern in the hotspot

4.2.3 Figures of Merit

Log-Likelihood: We use log-likelihood values to measure the goodness of fit of our models. The model with larger log-likelihood value is better than the one with smaller log-likelihood value [47, 182]. Therefore, for each probabilistic model we propose in the next sections, we compute the log-likelihood of that model based on specific training and test data sets. The log-likelihood of a model's probabilistic density function M with parameter Θ on a data set $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{N})$ is defined as:

$$\mathrm{LL}(\mathrm{M};\Theta;\mathrm{X}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \ln \mathrm{M}(\mathrm{x_i};\Theta)$$

This is a standard figure of merit in probabilistic learning. For the same data set, better fitting models have higher log-likelihood. Computing the log-likelihood of different models on the same training data provides an indicator of which model better fits the data that was used to select and train the models. Computing the log-likelihood of models given the test data provides an indicator of which model performs better on the same, yet unseen data i.e. provides a measure of how well the model captures the statistical variation in the training set.

Akaike Information Criterion: We also use Akaike information criterion (AIC) as an additional measure of the relative goodness of fit for our models and hence for model selection. AIC uses the concept of the information entropy and offers a relative measure of the information lost when a given model is used to describe real data [47, 183]. The general form for calculating AIC for a model with log-likelihood (LL) and the number of parameters K is defined as:

$$\mathrm{AIC} = -2 * \mathrm{LL} + 2 * \mathrm{K}$$

Given a set of candidate models for our data, the preferred model is the one with the minimum AIC value i.e. the model that minimizes the information loss and a penalty in increasing the number of estimated parameters.

4.3 Statistical Generative Models of 802.11 AP usage

4.3.1 Overview

In this section, we aim to provide descriptions of our proposed generative statistical models of access point (AP) usage and their mathematical formulations. We first consider models that assume the independence between consecutive daily event counts of APs i.e. time-independent models. Although this may not be the case, it caters to simpler models. Later in the section we will compare these models with others that do consider dependency between consecutive daily event counts i.e. time-dependent models.

Our time-independent models include a simple exponential distribution model, a discrete mixture of exponentials, a continuous Gamma mixture of exponentials, and an above-below model that allows presenting AP usage modeling problem in binary terms, for example if AP usage is high or low. Our time-dependent models include binary conditional probability (CPD) models and plugging-in above-below average models which provide a way to compare our time-independent models with time-dependent models. This is because our time-dependent models use binary variables for predicting next value for AP usage while the time-independent use event counts samples of APs. Moreover, we use seasoned ARIMA model (Auto-Regressive Integrated Moving Average) a standard tool for time series modeling, as an extra assessment of our time-dependent models. Experimental evaluation and validation results of all our proposed models in terms of log-likelihood and AIC values from specific pair of training/test are provided in section 4.4.

4.3.2 Exponential

We start our modeling endeavor by assuming that all daily event counts come from the same distribution, regardless of day or access point with an average rate λ. To explain this we pick a simple model i.e. exponential distribution because of its suitability in modeling events that occur continuously with constant average rate. We fit exponential model to our training and test data using maximum likelihood estimation function in Matlab. The following are the PDF and log-likelihood functions for this distribution on sample set $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{N})$:

$$\mathrm{p}_{\mathrm{exp}}(\mathrm{x};\ \lambda) = \lambda \exp(-\lambda \mathrm{x})$$

$$\mathrm{LL} = \ \mathrm{N} \ln \lambda - \lambda \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}$$

4.3.3 Discrete Mixture of Exponentials using K-Means

We next consider a mixture of distributions and in particular a discrete mixture of two or more exponential distributions (representing group of APs with different event counts averages) which has the following standard mixture PDF and log-likelihood functions on sample set $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{N})$:

$$\mathrm{p}_{\mathrm{dis\_mix\_exp}}(\mathrm{x}; \mathrm{W}, \Lambda) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{w}_\mathrm{k} \lambda_\mathrm{k} \exp(-\lambda_\mathrm{k} \mathrm{x})$$

$$\mathrm{LL} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \ln \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{w}_\mathrm{k} \lambda_\mathrm{k} \exp(-\lambda_\mathrm{k} \mathrm{x}_\mathrm{i})$$

Thus, each component of the mixture has a weight w_k, where $(\mathrm{w}_1, \mathrm{w}_2, \dots, \mathrm{w}_\mathrm{K}) = \mathrm{W}$ and $\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{w}_\mathrm{k} = 1$, and a parameter λ_k of the exponential, where $(\lambda_1, \lambda_2, \dots, \lambda_\mathrm{K}) = \ \Lambda$. The standard approach to generate a daily event count sample x using this model is straightforward: first choose a component k by generating a sample from a multinomial distribution with probabilities W. Then generate a value x from an exponential distribution with parameter λ_k.

We use Matlab's K-Means algorithm to cluster the averages of daily event counts of different APs in the training and test sets into K components, with K ranging between 2 and 20. The centers of the K clusters are used as the parameters of the Λ exponential components, whereas the percentages of APs assigned to the different components of the mixture were used as weights W.

4.3.4 Mixture of Exponentials with Gamma Scale

In a further attempt to gain better fitting in terms of log-likelihood and AIC values of the discrete mixture model of the previous subsection 4.3.3, without increasing complexity of the model (number of parameters), we employ a continuous mixture of exponentials with a parametric mixture model.

We chose Gamma distribution $\Gamma(\lambda; \alpha, \beta) = \beta^{\alpha}/\Gamma(\alpha).\lambda^{\alpha-1}\exp(-\beta\lambda)$ as the mixing model (also called scale) on the different values of λ. The mixture distribution on sample set $X = (x_1, x_2, \dots, x_N)$ is then:

$$p_{mix\_exp}(x, \lambda; \alpha, \beta) = \Gamma(\lambda; \alpha, \beta)\lambda \exp(-\lambda x)$$

$$p_{mix\_exp}(x; \alpha, \beta) = \int_0^{+\infty} \Gamma(\lambda; \alpha, \beta)\lambda_a \exp(-\lambda x)\, d\lambda = \alpha \frac{\beta^{\alpha}}{(\beta + x)^{\alpha+1}}$$

$$LL = N \ln\alpha + \alpha N \ln\beta - (\alpha + 1) \sum_{i=1}^{N} \ln(\beta + x_i)$$

The second form of p_{mix_exp} is the marginal distribution on x. For generating a sample with this model we first generate a λ from a Gamma distribution with parameters α and β, and then generate the sample from an exponential distribution with parameter λ.

4.3.5 Above-Below AP Daily Event Count Average

In the case where we want to think about AP usage as either high or low it may be useful to consider a binary variable θ that encodes whether a sample is above or below the average daily event count of its AP. This is so in order to represent APs with different usage activity levels in the 802.11 hotspot (see figure 6). Our proposed Above-below average model is better suited to explain situations where some group of APs in the hotspot are highly utilized i.e. high activity APs (above average APs) and those with low activity (below average APs). We see a clear separation between these groups of APs from the results of K-Means clustering (K=2) of the average number of events of APs, as shown in figure 6.

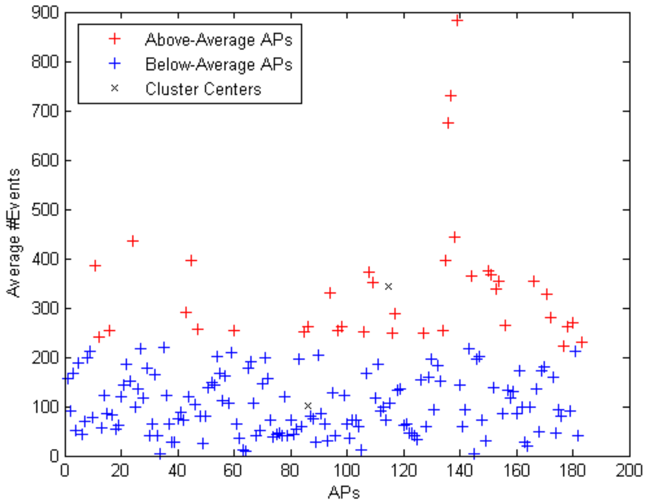


Figure 6. Clusters of APs based on above-below average event counts

For generating samples we use a three-level model. 1) For each daily event count sample we generate a value θ from a Gaussian distribution with support $[0,1]$ and parameters μ and σ. 2) We use θ as the parameter of a mixture of two components, each of which is a Gamma mixture of exponentials, with parameters α_1, β_1 for the component above the AP average and α_0, β_0 for the component below. 3) We finally generate a sample daily event count from an exponential distribution λ which we draw from the mixture of two components. This yields the following joint PDF, where $\Theta = (\mu, \sigma, \alpha_1, \beta_1, \alpha_0, \beta_0)$ is the parameter vector and $\mathcal{N}_{[0,1]}(\theta; \mu, \sigma)$ is a Gaussian distribution with support [0,1].

$$p_{\text{abv blw}}(x, \lambda, \theta; \Theta) = \mathcal{N}_{[0,1]}(\theta; \mu, \sigma)\ .$$
$$.\ \{\ \theta\ \Gamma(\lambda; \alpha_1, \beta_1)\ \lambda \exp(-\lambda x) +$$
$$+(1-\theta)\ \Gamma(\lambda; \alpha_0, \beta_0)\ \lambda \exp(-\lambda x)\ \}$$

Where

$$\mathcal{N}_{[0,1]}(\theta; \mu, \sigma) = \frac{\mathcal{N}(\theta; \mu, \sigma)}{\int_0^1 \mathcal{N}(\vartheta; \mu, \sigma) d\vartheta}, \theta \in [0,1]$$

The marginal distribution on x can be obtained by integrating the joint distribution over λ and θ. We first integrate on θ by computing $D(\mu, \sigma) = \int_0^1 \theta \mathcal{N}_{[0,1]}(\theta; \mu, \sigma) d\theta$ – the expected value of θ on $\mathcal{N}_{[0,1]}(\theta; \mu, \sigma)$ – using the standard error function erf(x). Then we integrate on λ using the result of the marginal distribution for the Gamma mixture of exponentials in the previous section. The marginal distribution on x is:

$$p_{\text{abv blw}}(x; \Theta) = D \frac{\alpha_1 {\beta_1}^{\alpha_1}}{(x + \beta_1)^{\alpha_1 + 1}} + (1 - D) \frac{\alpha_0 {\beta_0}^{\alpha_0}}{(x + \beta_0)^{\alpha_0 + 1}}$$

With

$$D(\mu, \sigma) = -\frac{\sigma}{\sqrt{2\pi}} \left(\exp\left(-\frac{(1-\mu)^2}{2\sigma^2} \right) - \exp\left(-\frac{\mu^2}{2\sigma^2} \right) \right) +$$
$$+ \frac{\mu}{2} \left(\operatorname{erf}\left(\frac{1-\mu}{\sqrt{2}\sigma} \right) + \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma} \right) \right)$$

The log-likelihood of this model on sample set $X = (x_1, x_2, \dots, x_N)$ is:

$$LL = \sum_{i=1}^{N} \ln\left(p_{\text{abv blw}}(x_i; \Theta)\right)$$

For parameter fitting and for computing the log-likelihood on the training and test data, we first split the training data into above-below AP average sets. For estimating $\mathcal{N}_{[0,1]}(\theta; \mu, \sigma)$ we calculate θ_a, i.e. the percentage of daily event counts above AP average for each AP, and fit a Gaussian distribution on the resulting set of θ_a to get an estimate of μ and σ. Similarly, we calculate the daily event count rates for samples above and below the AP average, λ_a^{abv} and λ_a^{blw}, and fit into two Gamma distributions resulting in the estimates for α_1, β_1 and α_0, β_0 respectively.

4.3.6 Binary Conditional Probability Models

Considering time dependencies usually increases the complexity of models. However, if we can represent data with e.g. binary variables, we can easily employ simple conditional probability distribution (CPD) table models. Entries in these tables give us the probability of observing a value (e.g. the value to predict t_T) given a set of other values $(t_1, t_2, \dots, t_{T-1})$. Because we are talking about binary variables, the number of entries in the table of a T-variable problem is 2^{T-1}, which is a manageable number if T is reasonably small.

Estimating a CPD table from the training and test data amounts to counting the number of occurrences of the 2^T different combination of values of variables $(t_1, t_2, \dots, t_{T-1}, t_T)$ and dividing by the sum of the occurrences of $(t_1, t_2, \dots t_{T-1}, 0)$ and $(t_1, t_2, \dots, t_{T-1}, 1)$.

Generating samples from this model implies defining the temporal order between variables $(t_1, t_2, \dots, t_{T-1}, t_T)$. Since in this chapter we are interested in daily event counts, our time unit is the day and $day(t_T)$ gives us the index of the day for which t_T occurs. An example of temporal order in these variables is $day(t_i) = day(t_{i-1}) - 1$ for all $i \in \{2, \dots, T\}$. We assume t_T occurs after all the other variables. Once all samples prior to t_T have been independently generated, we can use these samples as a lookup index in the 2^{T-1} table for the Bernoulli probability of the next sample.

A question that arises naturally is which variables $(t_1, t_2, \dots, t_{T-1})$ best help to predict t_T. Thus, here we look into a simple model in which $(t_1, t_2, \dots, t_{T-1})$ are the $T-1$ previous samples to t_T. We set T ranging from 2 to 12 and use a uniform prior to build 11 table CPDs for each AP. We then calculate average and standard deviation for each table entry. We calculate the log-likelihood of these models on the training and test sets by summing 1) the Bernoulli log-likelihoods of the $T-1$ first samples of each AP, setting parameter θ to the estimate of the average of $\mathcal{N}_{[0,1]}(\theta; \mu, \sigma)$ of the above-below model, with 2) the Bernoulli log-likelihoods of the remaining samples of each AP with the parameter obtained from our models - the value of the parameter changes with sample according to the previous samples and the table CPD. The Bernoulli log-likelihood of binary variable x and parameter θ is defined as:

$$\mathrm{LL} = \begin{cases} \ln\theta & : \mathrm{x} = 1 \\ \ln(1-\theta) & : \mathrm{x} = 0 \end{cases}$$

4.3.7 Plugging-in Above-Below Average Models

To compare time-dependent models which use binary above-below data sets with our time-independent models that use daily event counts, we use the above-below model as baseline and allow parameters μ and σ of the $\mathcal{N}_{[0,1]}(\theta; \mu, \sigma)$ distribution in the above-below model to change according to the CPD table values of the previous $T-1$ samples of the time-binary models.

For generating daily event counts using this mixed table CPD / above-below model, we start by generating $T - 1$ samples for each AP using the above-below model. Then, we iteratively lookup the μ_{CPD} and σ_{CPD} parameters in the CPD table that correspond to the previous $T - 1$ samples. We finally generate a daily event count sample from the above-below model with parameters μ_{CPD} and σ_{CPD}. We use the same log-likelihood function as the above-below event count average model in subsection 4.3.5, except that now each sample here can have a different set of μ and σ parameters.

4.3.8 ARIMA Model

We use ARIMA ((p, d, q)*(P, D, Q)$_h$) model [184] for modeling time series data in R package and fit into our training and test data sets. This provides an additional comparison point to our time-dependent models performance, in terms of log-likelihood and AIC values. The first three parameters of the ARIMA model i.e. (p, d, q) provide specification of the non-seasonal part of the ARIMA model: whereas *“p”* refers to the order of the autoregressive (AR), *“d”* the degree of differencing (I) and, *“q”* the moving average (MA) order of the model. The other three components *(P, D, Q)* provide specification of the seasonal parts of the ARIMA model: *“P”* refers to seasonal autoregressive (SAR), *“D”* number of seasonal differences and, *“Q”* seasonal moving average (SMA), plus seasonal period *“h”*.

We experiment with different combinations of parameters in the ARIMA model with *p, d, q, P, D and Q* $\in \{0,..,2\}$, while setting seasonal periodic component fixed at 7 (i.e. *h=7*), because we are using daily event count of APs and our period is one week (7 days). However, after looking at log-likelihoods and AIC values of different combinations of various seasonally differenced series, we picked ARIMA ((1, 0, 0)*(1, 0, 0)$_7$) model due to the minimum log-likelihood and AIC value. Our optimal ARIMA ((1, 0, 0)*(1, 0, 0)$_7$) model consists mainly of autoregressive part AR(1) and seasonal autoregressive part SAR(1) with seasonal periodicity=7; since the degree of differencing (I) and moving average (MA) terms are all zero together with their seasonality’s.

4.4 Experimental Evaluation of 802.11AP Usage Models

4.4.1 Exponential

The average daily event count obtained after fitting exponential distribution model to our training data is 132.53 keep-alive events. The log-likelihood values on the training and test sets are $LL_{train} = -4.3681\text{E}4$ and $LL_{test} = -1.2161\text{E}4$ respectively. Whereas their corresponding Akaike information criterion values are AIC_{train} = 8.7360E4 and AIC_{test} = 2.4324E4, respectively. These values are only useful when comparing with other models on the same data sets.

4.4.2 Discrete Mixture of Exponentials using K-Means

Figure 7 shows how the log-likelihood results changes with increasing number of components of the discrete mixture on the training and test data sets. Log-likelihood increases quickly for smaller number of clusters and appears not to increase as fast for larger number of clusters while the standard deviation of the log-likelihood for larger number of clusters is larger than for smaller number of clusters. This insight indicates that larger numbers of clusters provide better fitting to both training and test data. However, including more components in the discrete mixture increases the complexity of the model in terms of the number of parameters: for each additional component of the mixture we need an additional λ_k and w_k.

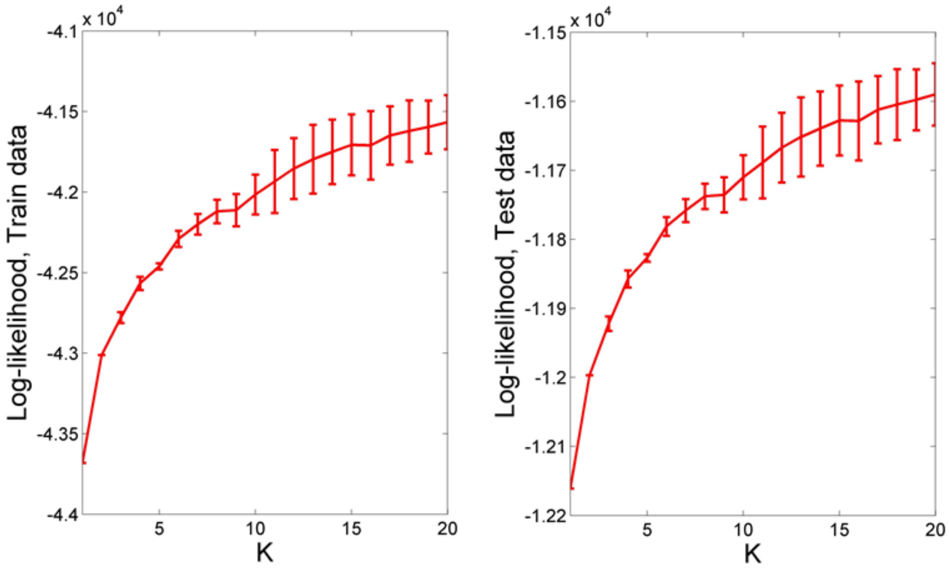


Figure 7. Log-likelihood of K-component exponential mixtures for the training set (left) and test set (right), and their 1-standard deviation error bars for 100 repetitions on each point. Notice that values for K=1 are equal to the results for the exponential model

4.4.3 Mixture of Exponentials with Gamma scale

We fitted Gamma distribution on the per-AP daily event count rate (i.e. one over the average) and we obtain the following results on the training and test sets: $\alpha = 0.46686$, $\beta = 9.5469$, for the log-likelihood: $LL_{train} = -4.2376\text{E}4$ and $LL_{test} = -1.1814\text{E}4$ and for the Akaike information criterion: AIC_{train} = 8.4756E4 and AIC_{test} = 2.3632E4. Comparing to results in figure 7, we see that the continuous mixture model outperforms K-Means discrete mixture models with K <= 5 (K = 5, AIC_{train} = 8.5018E4 and AIC_{test} = 2.374E4). This is an interesting result in terms of log-likelihood, Akaike information criterion, and model's complexity, as the continuous mixture model has only 2 parameters whereas the simplest discrete one that it outperforms has 9 (K=5, 2*K-1).

4.4.4 Above-Below AP Daily Event Count Average

After fitting this model to our training and test sets, we obtain the following parameter estimates: $\alpha_1 = 0.92572$, $\beta_1 = 8.1737\text{E}1$, $\alpha_0 = 0.23254$, $\beta_0 = 0.22538$, $\mu = 0.41563$, and $\sigma = 0.10866$. The log-likelihood values on the training and test sets using these parameters are: $LL_{train} = -3.9391E4$ and $LL_{test} = -1.1067\text{E}4$ and their corresponding Akaike information criterions are AIC_{train} = 7.8794E4 and AIC_{test} = 2.2146E4, respectively.

Comparing to figure 7, we see that our above-below model outperforms K-Means discrete mixture models with K <= 20, in terms of both log-likelihood and Akaike information criterion (*K =20, AIC_{train} =* 8.3078E4 and AIC_{test} = 2.3278E4*)*. In fact, this is true for all values of K that we tested. It also outperforms the exponential and Gamma mixture models. Although, we increased the complexity of the model by including 6 parameters rather than 2 (Gamma mixture of exponentials) and 1 (exponential), this is still remarkable as we outperform result of the model with 39 parameters (K-Means with K=20).

4.4.5 Binary Conditional Probability Models

Figure 8 shows the log-likelihood of the 11 different models for the binary above-below training and test data sets. The log-likelihood values increases consistently on the training set whereas on the test set it peaks at T = 8. This indicates that, whereas increasing T provides a better learning of the training data, at some point this is no longer beneficial for modeling the test data. Adding more AP's previous usage samples after T=8, only increases complexity of the model with insignificant gain in log-likelihood.

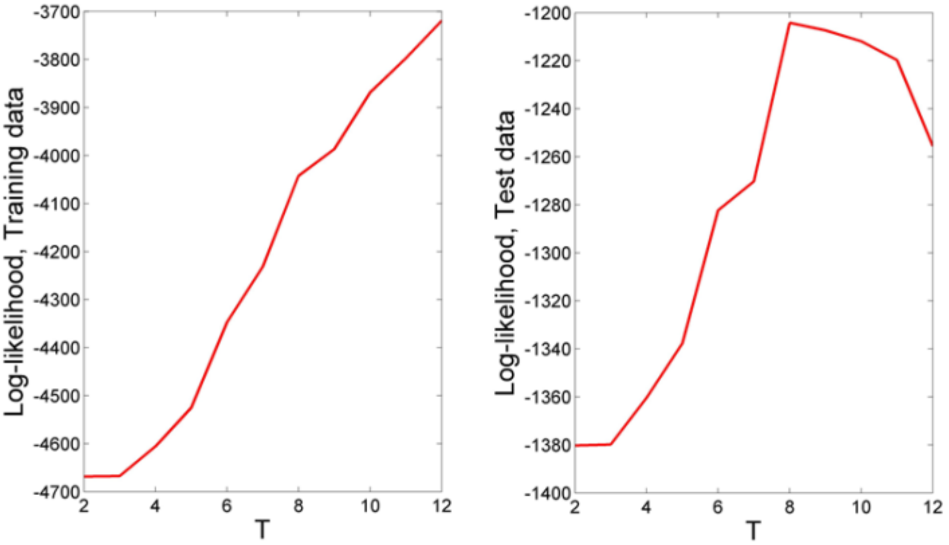


Figure 8. Log-likelihoods of the T=2:12 different models on the binary above-below training and test sets

4.4.6 Plugging-in Above-Below Average Models

Figure 9 shows the log-likelihoods of above-below models on the daily event count data sets. Notice that all these models have better log-likelihood values than the previous models, which did not consider time. However, we must also consider that each of these models requires the 6 parameters of the above-below model plus $2 * 2^{T-1}$ parameters for the table CPD. Figure 9 indicates $T = 6$ model with $6 + 2 * 2^5 = 38$ parameters is the best model in terms of log-likelihood for the training and test set amongst those presented previously with AIC_{train} = 7.8016E4 and AIC_{test} = 2.1986E4. This is still better in terms of number of parameters than the K=20 K-Means model with the Akaike information criterion AIC_{train} = 8.3078E4 and AIC_{test} = 2.3678E4.

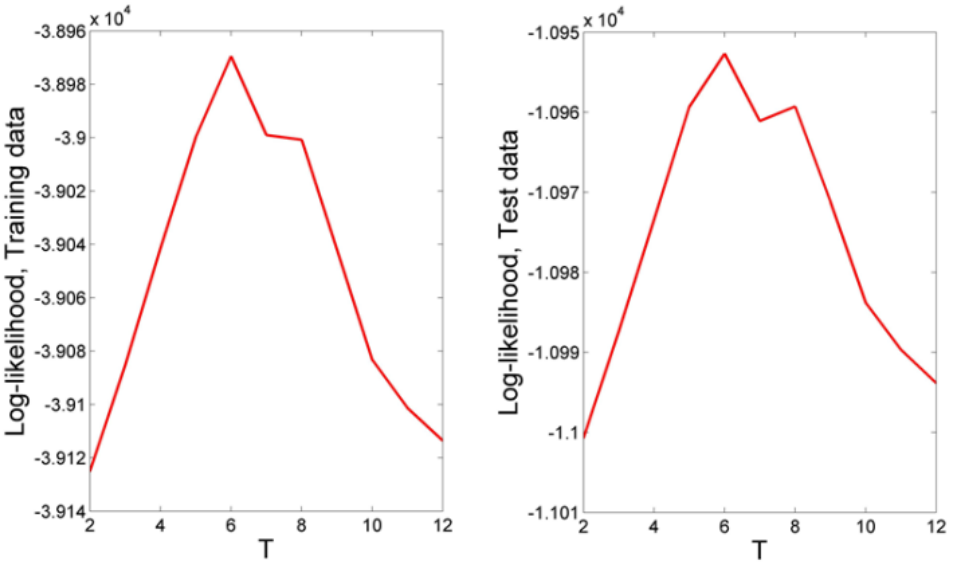


Figure 9. Log-likelihoods of the T=2:12 models on the daily event count training and test sets

4.4.7 ARIMA Model

Results after fitting our ARIMA ((1, 0, 0)*(1, 0, 0)$_7$) model to the training/test pair are given on table 2. However, these results can only be compared to the results of plugging-in above-below models of the previous subsection 4.4.6. This is because they were both computed based on daily event counts of AP, contrary to CPD models which use binary values. Comparing results to figure 9, we see that our plugging-in above-below models (T=2:12) outperform ARIMA model in terms of log-likelihood and AIC values on both training and test data sets. ARIMA is also outperformed by models in 4.4.1–4.44.

ARIMA Model	AR	SAR	Log-Likelihood	AIC
Training data	0.0926	0.13	-4.9679E4	9.9366E4
Test data	0.1519	0.0481	-1.3951E4	2.7910E4

Table 2. ARIMA Model fitting results

4.4.8 Cross-Validation

In order to gain further insight on the applicability of our models to different test and training data sets, we use a holdout cross-validation with the same number of training APs (143) and test APs (40). This setup is used throughout this chapter. 100 different partitions of the data into training and test data were randomly generated and all the models presented in this chapter were applied to these data partitions. Table 3 presents the average and standard deviation of our models' parameters. Someone trying to generate synthetic data out of these results should sample a Gaussian distribution with the two right columns as parameters, for each model parameter they would like to use.

Log-likelihood results depend on the data the models are applied to. With cross-validation, data sets change from partition to partition; and as such, the log-likelihood results cannot be compared directly. We calculate average, standard deviation, minimum, and maximum differences between the log-likelihood of the base exponential model and the other models, on the same data set for the 100 different training and test sets. These results are shown in figure 10 and generally confirm our pre-cross-validation results: 1) the Gamma mixture model on average outperforms K<=5 K-Means discrete exponential mixture models, although we did get at least an instance on which it performed worse on the test data than the exponential model; 2) the above-below model always outperforms Gamma and K<=20 K-Means discrete mixture models within one standard deviation; and 3) binary *T*-1 previous samples model on average outperforms the other models, with a peak at T=6.

Parameter Name	**Average**	**St.Dev.**
Exponential μ parameter	127.45	4.7637
Gamma mixture α parameter	0.52051	0.12000
Gamma mixture β parameter	12.646	7.7443
Above-below μ parameter	0.41566	4.5403E-3
Above-below σ parameter	0.10984	2.9368E-3
Above-below α_1 parameter	0.94505	0.10568
Above-below β_1 parameter	82.645	16.594
Above-below α_0 parameter	0.23709	5.7390E-2
Above-below β_0 parameter	0.30814	0.87809

Table 3. Model parameters

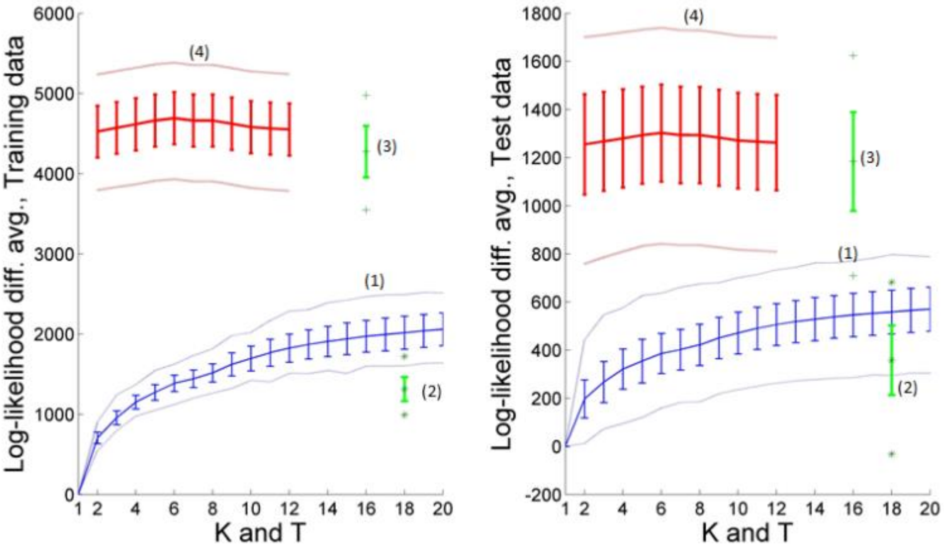


Figure 10. Log-likelihood differences from the base exponential model to: (1) K-Means discrete exponential mixture with K=1:20 (K=1 is the baseline exponential); (2) Gamma mixture at T&K=18;(3) above-below model at T&K=16; and (4) table CPD at T=2:12. We show average and standard deviation (solid lines), as well as minimum and maximum (dotted lines) of these differences on 100 random hold-out cross-validation training (left) and test (right) data sets

4.5 Time-Dependent Models Considering Different Settings of Variables

4.5.1 Overview

In this section, we look at a different arrangement of time variables contrary from that of section 4.3 and 4.4. Rather than predicting t_T based on all T-1 previous samples, here we consider different combinations of the T-1 previous samples, where one or more of the T-1 previous samples may or may not occur. The aim is to understand which combinations may result into better prediction of t_T. We limit T to 4 and use R=2 through 8 to index models with the following arrangement of variables *(t_{T-1}), (t_{T-2}), (t_{T-3}), (t_{T-1}, t_{T-2}), (t_{T-2}, t_{T-3}), (t_{T-1}, t_{T-3}),* and *(t_{T-1}, t_{T-2}, t_{T-3})*, as shown in figure 11.

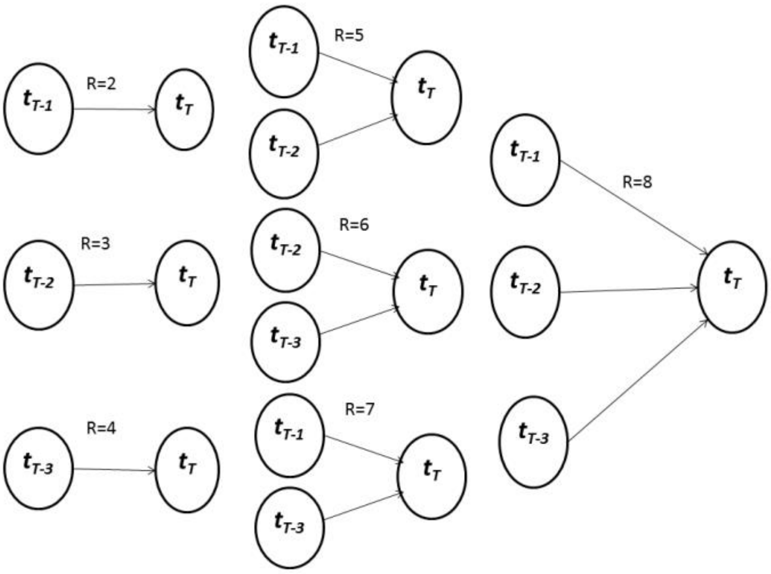


Figure 11. R=2:8 Different modeling proposals

4.5.2 Considering All-days AP Usage Samples

We first consider usage samples from all APs on all days to build table CPD and plugging-in above-below average event count models for R=2:8, in a similar fashion as in section 4.3.6. and 4.3.7. We then compute the log-likelihood and AIC for all R=2:8 models.

Cross validation results for table CPD models are shown in table 4. From table 4, it is apparent that the *LL* and *AIC* values on average are higher at R = 2, 5, 7, 8 and lower at R = 3, 4, 6. These results indicates using models (t_{T-2}) and (t_{T-3}) alone for table CPD models is of no benefit, while including model (t_{T-1}) in any settings result into better prediction of $t_{T.}$

Model	R=2	R=3	R=4	R=5	R=6	R=7	R=8
LL_{train}	-4518.07	-4812.89	-4818.67	-4486.86	-4799.51	-4488.81	-4415.89
LL_{test}	-1267.23	-1352.00	-1354.38	-1262.45	-1348.39	-1259.08	-1241.78
AIC_{train}	9044.13	9633.78	9645.33	8989.73	9615.02	8993.62	8863.78
AIC_{test}	2542.46	2711.99	2716.76	2540.89	2712.79	2534.17	2515.55

Table 4. Log-likelihood and AIC results of the R=2:8 CPD models on the training and test data sets

However, for the plugging-in above-below average models the log-likelihoods and AIC results change slightly across R=2:8 (see table 5) and in fact, on average are higher at R = 5, 6, 7, 8 and lower at R = 2, 3, 4. These results indicate including models (t_{T-1}), (t_{T-2}), and (t_{T-3}) alone is not beneficial in modeling both the training and test data for these event count models. However, the *LL* and *AIC* results in table 5 are not significantly different.

Model	R=2	R=3	R=4	R=5	R=6	R=7	R=8
LL_{train}	-33401.7	-33401.5	-33401.5	-33392.4	-33393	-33394.9	-33367.9
LL_{test}	-9405.87	-9405.90	-9405.99	-9398.72	-9398.7	-9398.7	-9390.72
AIC_{train}	66823.44	66823.03	66822.9	66812.8	66814.02	66817.71	66779.7
AIC_{test}	18831.73	18831.80	18831.97	18825.44	18825.4	18825.39	18825.44

Table 5. Log-likelihood and AIC results of the R=2:8 Plugging-in above-below average models

Moreover, when comparing results of table 5 and those of ARIMA model in table 2, it is evident that our plugging-in above-below models perform better on average in terms of log-likelihood and AIC values than ARIMA model fittings. This is true for all R=2:8 models on both training and test data sets. These results signify usefulness of the plugging-in above-below models in modeling event counts of AP over standard ARIMA model.

In this section, we studied the impact of time dependency ordering of AP's usage samples in the model. Our conclusion is two-fold: 1) for the binary CPD model the previous sample t_{T-1} is predominant and not including it has detrimental impact on the model ability to predict AP usage; 2) for the plugging-in above-below average model neither the order nor the number of previous samples significantly impacts model performance.

4.5.3 Considering Week Structure AP Usage Samples

Next, we study AP usage based on the structure of the week: 1) week-days usage, 2) week-ends usage, and 3) specific Monday through Sunday, seven individual days of the week usages. The aim is to derive usage models for each and to predict usage at APs, for example usage in the next week-day, next week-end, and in the next specific day given its previous *T*-1 usage samples.

We compute the log-likelihood and AIC values for all part of the week structure for both table CPD and plugging-in above-below average event count models. Table 6 shows cross validation results for all R=2:8 CPD models across all parts of week-structure. *LL* and *AIC* values are apparently higher at R = 2, 5, 7, 8 and lower at R= 3, 4, 6; revealing the same insights as CPD models of the previous subsection 4.5.2. That is using model *(t_{T-2})* and *(t_{T-3})* alone for CPD models is of no benefit for all parts of the week structure, while including model *(t_{T-1})* result into better prediction of t_T.

Model	LL/AIC	R=2	R=3	R=4	R=5	R=6	R=7	R=8
Week-Days	LL_{train}	-3002,85	-3077,58	-3151,44	-2922,43	-3048,75	-2970,57	-2904,14
	LL_{test}	-840,281	-861,025	-883,476	-818,418	-853,039	-831,285	-814,084
	AIC_{train}	6013,697	6163,157	6310,871	5860,862	6113,49	5957,141	5840,286
	AIC_{test}	1688,563	1730,049	1774,952	1652,836	1722,078	1678,57	1660,169
Week-Ends	LL_{train}	-370,016	-389,146	-396,024	-363,529	-383,359	-366,189	-358,573
	LL_{test}	-106,407	-112,057	-114,167	-105,009	-111,152	-105,938	-100,033
	AIC_{train}	748,0323	786,2912	800,0474	743,0589	782,7181	748,3776	749,1462
	AIC_{test}	220,8139	232,1136	236,3349	226,0189	238,3034	227,8765	228,0666
Individual-Days	LL_{train}	-2320,9	-2369,55	-2386,31	-2294,31	-2360,82	-2310,87	-2282,09
	LL_{test}	-666,731	-681,526	-686,413	-664,174	-682,578	-667,475	-666,944
	AIC_{train}	4649,806	4747,093	4780,625	4604,622	4737,63	4637,733	4596,187
	AIC_{test}	1341,463	1371,052	1380,826	1344,348	1381,156	1350,949	1365,888

Table 6. Log-likelihood and AIC results of the R=2:8 CPD models across all parts of the week-structure

Nevertheless, for fitting plugging-in above-below average models on daily event counts of APs, we allow parameters μ and σ of the $N_{[0,1]}$ (θ, μ, σ) distribution in the model to change according to either week-days, week-ends, or individual days R=2:8 CPD table values. Consequently, we compute the log-likelihood and AIC values across all part of the week structure. Cross-validation results in table 7 indicates that log-likelihood and AIC values change very slightly across R=2:8 and are apparently higher at R = 5, 6, 7, 8 and lower at R= 2, 3, 4; this is true for all parts of the week structure. These results indicate including more previous samples *(t_{T-1})*, *(t_{T-2})*, and *(t_{T-3})* is of no significant benefit in modeling both training and test data for these models.

Model	LL/AIC	R=2	R=3	R=4	R=5	R=6	R=7	R=8
Week-Days	LL_{train}	-28428,5	-28431,5	-28427,5	-28423,1	-28422,9	-28423,1	-28415,3
	LL_{test}	-8022,01	-8022,02	-8022,02	-8021,92	-8021,99	-8021,89	-8021,81
	AIC_{train}	56876,92	56883,0	56875,01	56874,19	56873,81	56874,23	56874,62
	AIC_{test}	16064,02	16064,03	16064,04	16071,83	16071,97	16071,78	16087,61
Week-Ends	LL_{train}	-3838,84	-3835,07	-3832,96	-3827,36	-3822,69	-3824,84	-3820,1
	LL_{test}	-1028,55	-1035,22	-1028,93	-1023,39	-1024,19	-1024,39	-1015,8
	AIC_{train}	7697,689	7690,143	7685,919	7682,71	7673,377	7677,677	7684,195
	AIC_{test}	2077,107	2090,441	2077,854	2074,77	2076,376	2076,784	2075,591
Individual-Days	LL_{train}	-21178,7	-21189,9	-21186,6	-21159,5	-21157	-21154,2	-21162,9
	LL_{test}	-5941,63	-5940,76	-5941,33	-5924,46	-5922,94	-5923,38	-5918,36
	AIC_{train}	42377,45	42399,71	42393,13	42347,03	42342,05	42336,34	42369,82
	AIC_{test}	11903,26	11901,51	11902,65	11876,92	11873,87	11874,76	11880,71

Table 7. Log-likelihood and AIC results of the R=2:8 Plugging-in models across all parts of the week-structure

Moreover, the log-likelihood and AIC results after fitting ARIMA model on the week structure AP usage samples are shown on table 8. Results indicate that the plugging-in above-below models based on week-days, week-ends, and individual days AP usage samples in table 7 outperformed ARIMA model fittings in terms of both log-likelihood and AIC values. This is true for all R=2:8 on both training and test data, across all parts of the week structure.

ARIMA Model	AR	SAR	Log-likelihood	AIC
Week-Days Train	0,0688	-0,0324	-34526,7	69061,33
Week-Days Test	-0,0179	0,0664	-11072,0	22151,92
Week-Ends Train	0,1508	-0,0214	-12456,5	24921,06
Week-Ends Test	0,113	0,068	-4361,19	8730,38
Individual Days Train	0,6022	-0,2566	-46869,7	93795,43
Individual Days Test	0,073	0,4028	-15382,8	30821,61

Table 8. ARIMA model fitting results for all parts of the week-structure

4.5.4 Comparing All-days Model Vs. Hybrid Model

In this section, we compare the all-days model of subsection 4.5.2 to a hybrid model of the week-days and week-ends models. Depending on the day of the week, the hybrid model will choose to apply the week-days model (if it's a week day) or the week-ends model (if it's Saturday or Sunday). We set the all-days model as a baseline and monitor the gain in log-likelihood incurred using the hybrid model.

Cross-validation results indicate that the log-likelihood and AIC values of the hybrid CPD and plugging-in above-below models on average are higher than that of the all-days model for all R=2:8 (see table 9). This means the hybrid model is better than the all-days model in the prediction of t_T for all possible values of R. However, using the hybrid model increases model complexity by order two.

Model	LL/AIC	R=2	R=3	R=4	R=5	R=6	R=7	R=8
CPD	LL_{train}	-1145,20	-1346,16	-1271,21	-1200,90	-1367,40	-1152,05	-1153,18
	LL_{test}	-320,54	-378,92	-356,74	-339,02	-384,20	-321,86	-327,66
	AIC_{train}	2282,40	2684,33	2534,41	2385,81	2718,81	2288,10	2274,35
	AIC_{test}	633,08	749,83	705,47	662,04	752,41	627,72	627,31
Plugging-In	LL_{train}	-1134,36	-1134,93	-1141,04	-1141,94	-1147,41	-1146,96	-1132,50
	LL_{test}	-355,31	-348,66	-355,04	-353,41	-352,52	-352,42	-353,11
	AIC_{train}	2248,83	2249,89	2261,97	2255,90	2266,83	2265,80	2220,88
	AIC_{test}	690,60	677,33	690,08	678,84	677,05	676,83	662,24

Table 9. The gain in log-likelihood and AIC values across R=2:8 of the hybrid model over all-days model

4.5.5 Comparing All-days Model Vs. Individual Days Model

We lastly compare the all-days model of section 4.5.2 to the individual day's usage model. As such, depending on the day of the week, the individual day's model will choose to apply appropriate model for a given particular day of the week e.g. if it is Monday an AP usage models for Monday will be chosen. We set the all-days model as a baseline and monitor the gain in log-likelihood incurred using the individual day's model.

The log-likelihood and AIC values of the individual days CPD and plugging-in above-below model are significantly higher than that of the all-days model (table 10). This shows that the individual day's model is considerably better than the all-days model in the prediction of t_T . This insight is confirmed by cross-validation results on training and tests data sets, see table 10. However, using the individual day's model increases model complexity by order 7.

Model	LL/AIC	R=2	R=3	R=4	R=5	R=6	R=7	R=8
CPD	LL_{train}	-2197,17	-2443,34	-2432,36	-2192,55	-2438,69	-2177,94	-2133,80
	LL_{test}	-600,50	-670,47	-667,97	-598,28	-665,81	-591,61	-574,84
	AIC_{train}	4394,33	4886,69	4864,71	4385,11	4877,39	4355,88	4267,59
	AIC_{test}	1201,00	1340,94	1335,93	1196,55	1331,63	1183,22	1149,66
Plugging-In	LL_{train}	-12223,0	-12211,6	-12214,9	-12232,9	-12236,0	-12240,7	-12205,0
	LL_{test}	-3464,24	-3465,14	-3464,66	-3474,26	-3475,76	-3475,32	-3472,36
	AIC_{train}	24445,99	24423,32	24429,77	24465,77	24471,97	24481,37	24409,88
	AIC_{test}	6928,47	6930,29	6929,32	6948,52	6951,53	6950,63	6944,73

Table 10. The gain in log-likelihood and AIC values across R=2:8 of the individual days model over the all-days model

4.5.6 Summary

In this section we consider all-days and week structure usage sample of APs. We compared different binary CPD models and plugging-in above-below average models based on all-days and on different parts of the week structure AP usage samples, with different time dependency ordering of samples. In addition to confirming our guess that considering week structure has a positive impact on the performance of all models, our conclusions are: 1) individual days model has approximately twice performance gain from all-days model when compared to hybrid model on CPD data; whereas 2) on event count data using the plugging-in above-below average model, this performance gain is an order of magnitude higher than that of CPD model; 3) binary CPD performance continues to show dependency on previous sample and plugging-in above-below average model performance on test data continues to be independent of AP's previous sample ordering, regardless of week-day, week-end, hybrid, or individual day model and data.

4.6 Performance Comparison of AP Usage Models based on Different Data Sets

4.6.1 Overview

In this section, we want to understand the applicability and feasibility of using the binary conditional probability models and plugging-in above-below models of the previous section 4.5 when applied to different usage data. Namely, we want first: 1) to employ model derived from the parameters of **all-days model** and apply it to week-days and week-ends AP's usage data; next 2) to employ model derived from the parameters of **week-days model** and apply it to all-days and weekends AP's usage data, and lastly; 3) to employ model derived from the parameters of the **week-ends model** and apply it to all-days and week-ends AP's usage data. Intuitively, models should have better performance on the data set on which they were trained, e.g. the all-days CPD or plugging-in above-below model should have better log-likelihood on all-days data than on week-days or week-ends AP usage samples. The aim of this section is to understand the extent to which this reasoning is correct.

4.6.2 Binary Conditional Probability Models

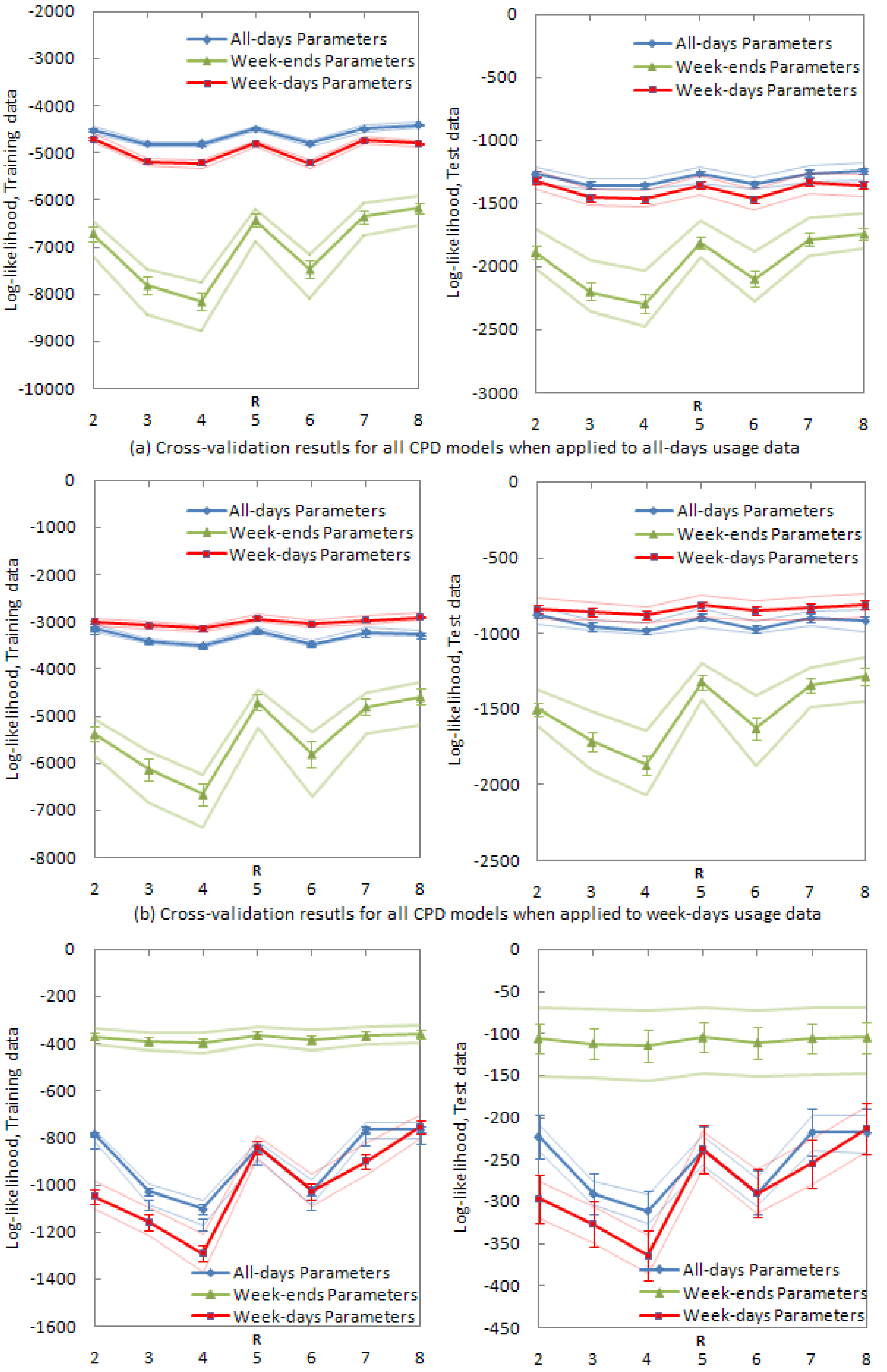
To realize the above set objectives, we first consider binary R=2:8 CPD models. Figure 12 shows cross-validation results for these CPD models based on the aforementioned 1-3 scenarios. On average: 1) the binary CPD models based on all-days parameters outperformed others when applied to the all-days AP's usage data; similarly 2) the binary CPD models based on week-days parameters outperformed others when applied to the week-days AP's usage data; and likewise 3) the binary CPD based on week-ends parameters outperformed others when applied to the week-ends AP's usage data. This is true on both training and test data sets across all R=2:8; hence agrees with our first intuition that models should perform better on the data set to which they were trained.

4.6.3 Plugging-in Above-Below Average Models

We also consider R=2:8 plugging-in above-below event count models. Figure 13 shows cross-validation results for these event count models given the same conditions stated in 4.6.1. The log-likelihood results change slightly across R. On average: 1) the log-likelihood of the plugging-in above-below average models based on the parameters of all-days given the all-days AP's usage data outperformed others on training data set, while for the test data set plugging-in models based on week-ends usage data perform better than others; 2) the log-likelihood of the plugging-in above-below average models based on week-days parameters given the week-days usage data outperformed others on the training data set, while for the test data plugging-in week-ends models perform better; 3) the log-likelihood of the plugging-in above-below average models based on the parameters of all-days to the week-ends usage data outperformed others on both training and test data sets. However, the maximum, minimum, and standard deviation in the cross-validation results continue to indicate large dispersion throughout R=2:8.

4.6.4 Summary

In this section we tried to understand how different week structure models perform when applied to parts of the week AP's usage data for which they were not trained on. For binary CPD models we confirmed our intuition that e.g. week-days model had better performance on week-days usage data than on all-days or week-ends usage data; this is also true for all other CPD models. However, for the event count data using the plugging-in above-below average models we could not confirm this, given the large dispersion of the results.

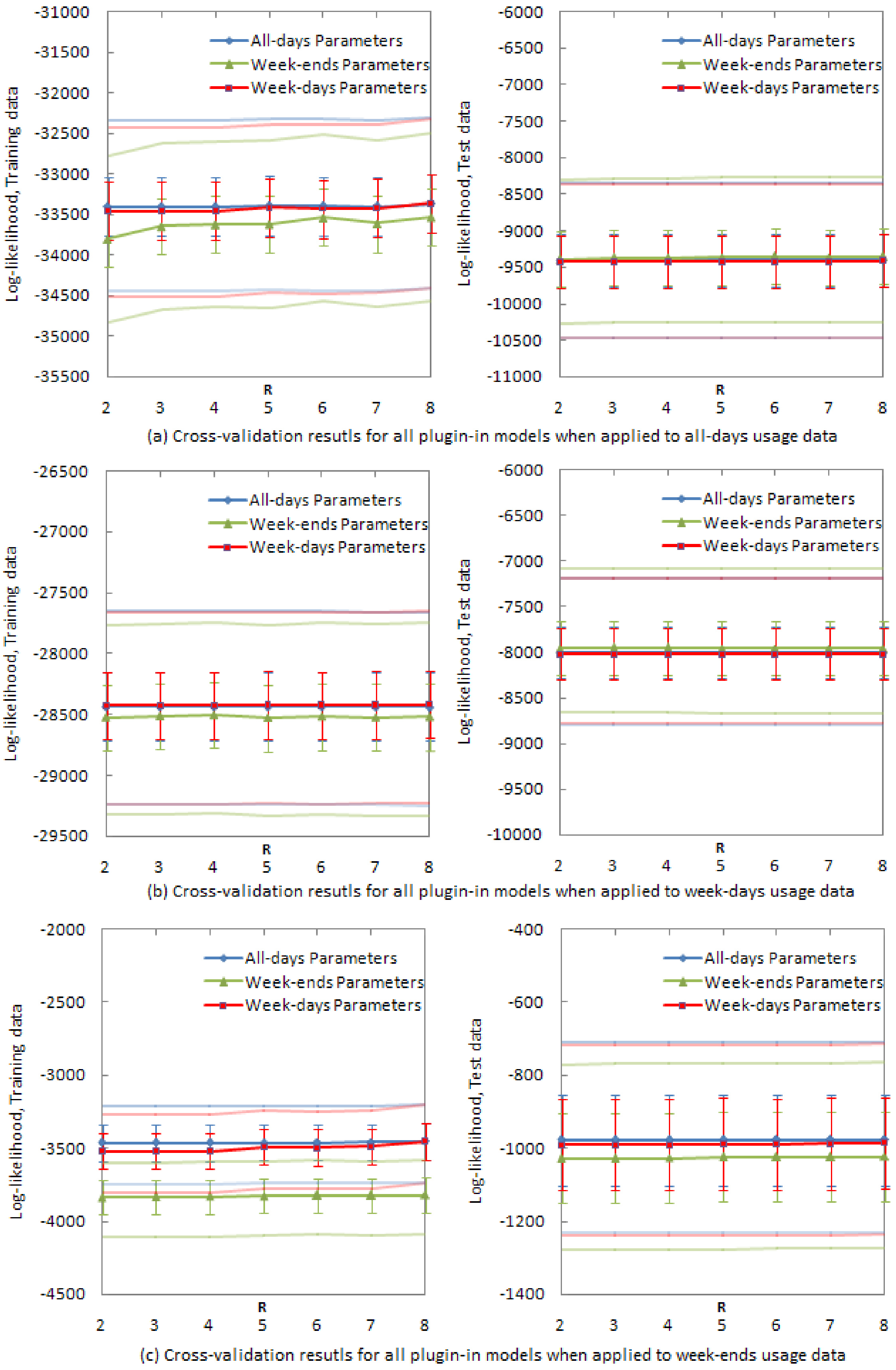


(a) Cross-validation resutls for all CPD models when applied to all-days usage data

(b) Cross-validation resutls for all CPD models when applied to week-days usage data

(c) Cross-validation resutls for all CPD models when applied to wek-ends usage data

Figure 12. 100 random hold-out cross-validation results of the R=2:8 CPD models when applied to different AP usage data, average (solid lines), standard deviation (bars), minimum and maximum (dotted lines)



(a) Cross-validation resutls for all plugin-in models when applied to all-days usage data

(b) Cross-validation resutls for all plugin-in models when applied to week-days usage data

(c) Cross-validation resutls for all plugin-in models when applied to week-ends usage data

Figure 13. 100 random hold-out cross-validation results of the R=2:8 Plugging-in above-below average models when applied to different usage data, average (solid lines), standard deviation (bars), minimum and maximum (dotted lines)

4.7 Conclusions

In this chapter, we presented an approach to model 802.11 AP usage that focuses on daily keep-alive event counts proportional to the time users are connected to an AP, and used generative probabilistic models such as a Gamma mixture of exponentials, binary Conditional probability models, and a plugging-in above-below AP average model that makes it easier to consider dependencies between consecutive samples in time. We compared our models with the log-likelihood and AIC values standard figures of merit in statistical learning, based on specific training and test data sets. We conclude that the increase in complexity of the T = 6 plugging-in above-below models with 38 parameters (that has the best log-likelihood and AIC values on both training and test sets) leads on average to a much smaller gain, in terms of both log-likelihood and AIC values, as compared to complexity incurred by using the Gamma distribution with 2 parameters and the above-below model with 6 parameters. These conclusions are supported by a 100-fold random holdout cross-validation. Cross-validation results also confirmed the significant gain in log-likelihood and AIC values on the training and test data sets of using 1) a hybrid model for week-days/week-ends and 2) individual day's model for a given specific day of the week, as well as 3) the meager impact of adding more time dependency through AP previous samples and of changing time variable settings for the event count plugging-in models.

We believe models presented in this chapter can be useful as a complement to the existing performance management techniques of large-scale 802.11 networks. Based on the insights derived from our models, particularly conditional probability distribution (CPD) models, 802.11 administrators might be able to understand and predict AP usage on daily and weekly basis, and thus allocate network resources according to expected usage, for example, installing additional APs on days where is usage is known to be high. Moreover, anomaly detection of 802.11 AP usages can also be enhanced based on the insights derived from our models e.g. baselines or usage levels of different APs and days can be established. This means that high usage (e.g. event counts) for a particular AP that is not known usually to be very active could indicate the possibility of an anomaly such as AP overload. The absence of activity/events in an AP that is usually known for its high usage could indicate an anomaly such as an AP crash. The same idea can be applied for detecting usage anomalies on different days of the week, with exception for holidays. For example, days which are usually known for low usage (event counts), high usage in these days could mean an unusual event has happened in the infrastructure (e.g. conference), and vice versa. In this thesis, we do not further explore our AP usage models for anomaly detection. We focus instead on a pattern that we call "Abrupt Ending" of 802.11 AP connections. In the next subsequent chapters, we will explain methods, algorithm, and models for detecting and characterizing such pattern of anomalies in AP usage based on the collected 802.11 usage data.

Chapter 5

Anomaly Detection of 802.11 AP Usage - Abrupt Ending of 802.11 AP Connections

5.1 Introduction

In this chapter, we introduce our first anomaly detection work. Our emphasis is on detecting patterns of anomaly in the usage of 802.11 APs. In our context, an anomaly is a pattern that does not conform to normal 802.11 AP usage behavioral patterns. Anomaly detection refers to the process of finding these non-conforming patterns. The importance of anomaly detection in 802.11 networks or in any other domain is that usually anomalies translates to significant, and often critical, actionable decisions in an attempt to remedy the anomalous situation.

In this chapter we focus on a usage pattern named “abrupt ending” of 802.11 AP connections that happens when a large number of user sessions in the same access point (AP) end within a one second window. Our goal is three-fold: 1) to investigate a method for detecting abrupt ending of AP connections from massive 802.11 usage data, 2) to examine possible causes that influence AP abrupt ending occurrence; and 3) to propose models for proper characterization of their occurrences. The emphasis on detecting and characterizing patterns of anomaly in AP usage is critical for fault management of large-scale 802.11 networks as it may help network administrators to quickly and efficiently detect and fix different connectivity problems that users of large-scale 802.11 networks often face, for example authentication failure and intermittent connectivity of user connections. We will follow this up in the chapter 6.

We proceed as follows in this chapter. In section 5.2, we define abrupt ending of AP connections, and describe what happens upon their occurrences, including resulting connectivity disruptions to wireless users. We also investigate a method for detecting abrupt ending of AP connections from the collected massive 802.11 usage data in this section. In section 5.3, we examine causes of AP abrupt endings manifestation from different perspectives including users, user devices, AP models, AP locations, and aggregate usage of the 802.11 infrastructure network. In section 5.4, we propose statistical models for characterizing occurrence of abrupt endings with respect to aggregate usage of 802.11 networks, in terms of total number of sessions. In section 5.5 we present concluding remarks.

5.2 Abrupt Ending Characterization

5.2.1 Definition

Abrupt ending of AP connections refers to a situation where an 802.11 AP drops all or a significant part of its users' connections within a one second time window. During abrupt endings, mobile stations typically change association to other APs and it can take few seconds, minutes, or even hours before an AP starts accepting new connections again. Usually, user connections are established after the exchange of authentication and association frames between user device's NIC and target AP, and then followed by DHCP packets exchange between user mobile device and a DHCP server before an IP address is assigned to the user to start using the 802.11 network [1].

To change AP association mobile stations must first probe APs in their vicinity before performing handoff. Each time a handoff happens, usually management frames are exchanged between the mobile station and the target AP, thus keeping the wireless medium busy and preventing other mobile stations from accessing the wireless medium [119, 123, 155]. The need to associate and re-associate particularly during 802.11 AP abrupt ending of connections can have significant impact on the users-AP connectivity and on the performance of users' traffic. For instance, consider an AP that stops accepting connections for five minutes after encountering abrupt ending; this will create massive connection disruptions not only to users associated to this AP, but also to many users in the vicinity of this AP. Mobile stations associated to this AP will repeatedly perform probing and try to associate with various APs in their vicinity, in most cases without being able to create stable/reliable connections. This connectivity disruption can potentially spread across many parts of the network, if not controlled.

Since the association mechanism employed by 802.11 devices is typically based on highest received signal strength, and once the AP which provides highest signal strength is momentarily down, the possibility of stations establishing stable connection to other APs is very low, resulting in inconsistent connectivity. This intermittent connectivity can lead to even more control and management traffic, which can further degrade users-AP connectivity and performance of their traffic. This is similar to what is reported in previous work [123] where an entire 802.11 WLAN collapsed as it could not sustain the heavy control, management, and data packet processing required by arrival of high concentration of users to its APs, each time APs started accepting connections. This is also one of the possible scenarios that could emerge upon abrupt ending of AP connections. Because mobile stations typically tend to probe to find better connection, it can possibly happen that a large group of users discovers one AP with high signal strength and as a result all users migrate and associate to such AP. If this AP already has too many users associated, then the possibility of AP halt or crash is strong. We will further address resulting patterns of AP abrupt ending of connection such as AP overload and AP crash in chapter 6.

5.2.2 Trace Data Characteristics

We use RADIUS authentication data collected at the hotspot of the Faculty of Engineering of the University of Porto (FEUP). The trace data was captured from November 2, 2006 to March 27, 2009, and consists of 802.11 mobile stations to AP association log records stored at a RADIUS authentication server for 878 consecutive days. Each log event recorded (i.e. START, ALIVE, or STOP) includes timestamp, session ID, association duration, number of input and output bytes and packets among the other attributes [36]. Table 11 depicts overall hotspot usage characteristics across academic semesters as observed from our collected 802.11 traces.

Moreover, our trace data reveal that there about 207 APs covering the campus over the two and a half years. The number of unique users observed during this trace period is over 14 thousand and the corresponding number of sessions (START-STOP) established by these users is close to 6 million. In general, the hotspot experienced an increase in total network usage over time (and consequently abrupt endings). The exception is the second semester of academic year 2008/2009 which does not include all data of the entire academic semester.

Academic Semester	# Abrupt Endings	# APs	# Users	# Sessions	Total #Input Bytes (TB)	Total #Output Byes (TB)
S1 06/07	91	146	2861	751377	4.22996	8.34779
S2 06/07	99	157	3633	1282482	11.7742	18.6249
S1 07/08	406	161	4512	1349334	7.80847	23.0022
S2 07/08	2338	195	5202	1271549	10.5128	25.8237
S1 08/09	3646	200	8260	1275978	10.484	32.4974
S2 08/09	2693	207	7400	1062869	2.3275	7.1315

Table 11. High-level evolution of the hotspot usage over the 2 and a half year

5.2.3 Threshold for Abrupt Ending of AP Connections

In this section, we describe how to detect abrupt ending of AP connections from the collected 802.11 usage data. Typically, the RADIUS server records log events with a granularity of one second according to the RADIUS protocol specifications [35, 36]. Taking this fact into account, we then count how many sessions from the same AP end at the same second to get a sense of simultaneousness of session endings. We define “n-endings second” as a second where n sessions from the same AP end. We define “n-ending session” as a session that ends in an n-endings second. From our data sets we find the bulk of the sessions (99.45%) end at 1-ending second. When looking for an n-endings seconds with n > 1 we find a sharp drop in the number of session endings and a smaller number of APs where these session endings are observed (table 12). We define an abrupt ending of 802.11 AP connections as an n-ending second with n >= 3, i.e. where three or more session end in one second window at the same AP; signifying an abnormal sessions ending to AP.

n	% sessions in n-ending seconds	AP Count
1	99.45%	207
2	0.42%	197
3	0.025%	117
4	0.014%	67
5	0.0085%	48
6	0.0082%	41
7	0.0075%	35
8	0.0067%	34
9	0.0064%	31
10	0.0062%	30

Table 12. Percentage of n-ending sessions and their respective AP counts

Table 12 also shows that the decrease in the number of n-ending sessions is long-tailed rather than exponentially distributed. This suggests to some extent that session endings are not independent from each other as we may have assumed. We consider the following arguments when discussing why session endings are not independent. 1) Usage may not be independent. It is reasonable to expect correlated user behavior in an academic campus. For example: when class is over, students switch-off their laptops at approximately the same time. However, this correlated user behavior is unlikely to map to correlation in session endings at a one second granularity. Even if users switch-off their laptops at exactly the same time, some operating systems and hardware will take longer to disassociate from the AP than others. This will spread session endings across different seconds. 2) Wireless interference can affect multiple mobile stations simultaneously. 3) Configuration issues or software bugs in APs and in the network and network services can affect all users at once in one AP. In chapter 6 we will identify several anomaly-related patterns in the data set that could be related to these and other causes of 802.11 AP abrupt endings.

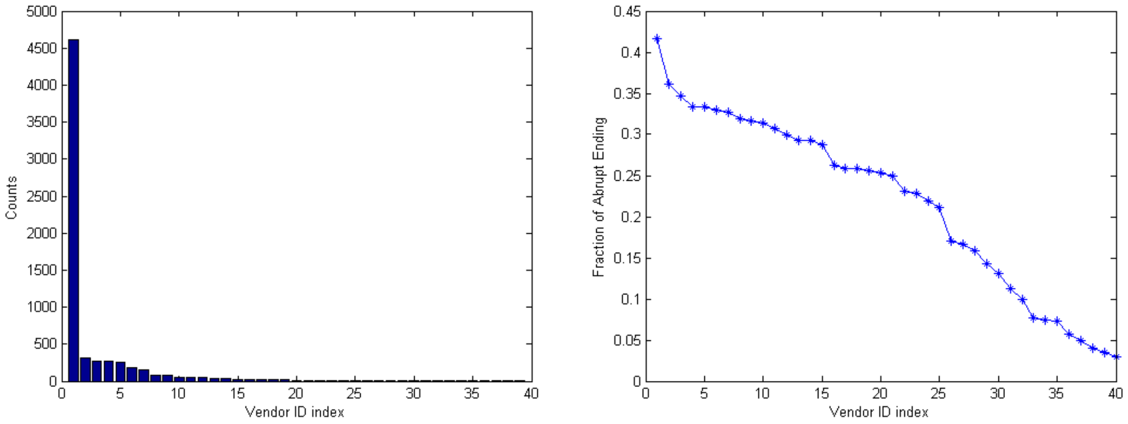
5.2.4 Abrupt Ending Data Overview

We observe 127 APs out of total 207 APs experienced at least one abrupt ending throughout the trace period. 47 APs encountered more than 100 abrupt endings while 80 APs encountered less than 100 abrupt endings. However, one AP in particular experienced over 1000 abrupt endings during the entire trace period. Looking at daily statistics, we observe 595 days out of total 878 days with at least one abrupt ending. The maximum observed is 65 abrupt endings in one day. The remaining 283 days were without any abrupt ending; these are days where the 802.11 infrastructure was less utilized (week-ends and holidays). We will use this abrupt endings data set to study several relationships that may exist related to their causality in the subsequent section 5.3. We will also use this data set as input to our proposed pattern detection algorithm in chapter 6 in order to detect and characterize different anomaly-related patterns of AP usage such as AP interference, AP overload, and AP crash.

5.3 Possible Causes

5.3.1 Abrupt Ending and User's Devices

We use MAC addresses to classify user's devices by vendor, aiming to investigate if devices from some specific vendors are more prone to 802.11 AP abrupt ending of connections or not. We identified total of 140 unique vendors, of which 100 (71%) had been using the 802.11 infrastructure rarely i.e. they established less than total of 10 sessions throughout the trace period. We ignored them in the analysis. Figure 14(a) depicts popularity of each vendor in terms of total device counts. The fraction of total abrupt ending sessions to the total number of sessions established by each vendor's device is presented by figure 14(b). The higher fractions observed in this plot results from popular vendors. This also means devices that establish more sessions in the 802.11 networks are more likely to encounter abrupt ending than those which appear to establish fewer sessions. The top ten devices observed in our 802.11 infrastructure and their percentages of abrupt ending sessions are as follows: Intel 41%, Apple 36%, Askey 34%, ASUSTec 33%, Sitecom 33%, AzureWave 32%, LITE-ON Te 31%, Hon Hai 30%, Nokia 29%, and Samsung 28%.



(a) Device counts for each vendor

(b) Fraction of total abrupt ending sessions to the total number of sessions for each vendor

Figure 14. Popularity of vendors and abrupt ending occurrences

5.3.2 Abrupt Ending and AP Models

Here we investigate abrupt ending occurrences versus AP models. When analyzing abrupt ending occurrences against AP model, we aim to understand if some AP models are more prone to abrupt ending than others. In the trace data we observe one particular model (Cisco C1250) to have been involved predominantly in abrupt endings. All higher counts (150-1007 abrupt endings per AP) resulted from this particular model even though they were not largely deployed in the campus 802.11 infrastructure. This means abrupt ending of connections can be caused by configuration issues of 802.11 APs and bugs in the APs. The hotspot has 5 models of Cisco APs, their total number in the network and number of abrupt ending encountered together with the fraction of abrupt ending sessions to the total sessions are shown in table 13. Cisco C1250 APs appear to have higher counts of abrupt endings sessions in relation to the total number of sessions established than other AP models, hence indicating possibility of faults.

AP Model (Cisco)	Total Number (AP)	#Abrupt Ending (AE)	Total #Sessions	Total #AE Sessions	Fraction AE Sessions
C1100	80	178	1889778	680	0.000359
C1130	33	525	466701	2402	0.005146
C1140	47	32	33552	131	0.003904
C1200	27	527	881817	2059	0.002334
C1250	20	6515	1880799	65310	0.034724

Table 13. AP models and their associated counts and fraction of abrupt endings

5.3.3 Abrupt Ending and User IDs

Similarly, in this section we want to investigate if there is any possibility that users are responsible for AP abrupt endings. In particular, we are interested to see if the presence of some users can result in abrupt ending of connections to AP. Trace data indicates about 54% of users have encountered abrupt endings at least once; we observed a maximum of 500 abrupt endings for one specific user. The fraction of abrupt ending sessions to the total number of sessions established by each user is depicted by its CDF in figure 15. Figure 15 indicates there is a pronounced tail to the distribution, showing the presence of a small number of users who had majority of their sessions involved in abrupt endings.

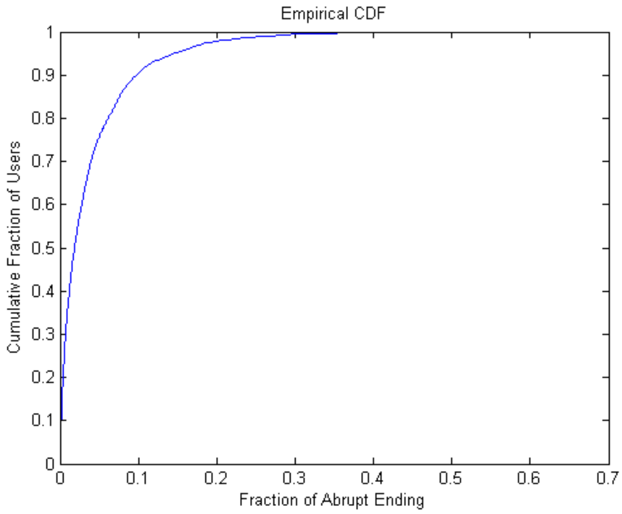


Figure 15. Fraction of abrupt ending sessions to the total number of sessions per user

The CDF plot indicates 80% of users have less than 10% of their sessions involved in abrupt endings. On average user sessions encountered 0.0383 abrupt endings. The maximum fraction of session's encountered abrupt ending by a user is 0.6970. Nevertheless, there is no clear evidence implicating a subset of users to 802.11 AP abrupt ending of connections. The higher fractions observed in figure 15 have more to do with locations users tend to associate and use the 802.11 network, e.g. library and class rooms, rather than what users had been doing in the network in terms of usage.

5.3.4 Abrupt Ending and AP Locations

We then seek to investigate the influence of AP locations on abrupt ending of connections. The aim is to understand the impact of different AP locations on abrupt ending of connections. In our 802.11 network, APs were assigned names based on the locations on which they provide 802.11 coverage. Information about the location of a subset of these APs is known to us through the "WISPRLOCATION" field of the RADIUS authentication data set [36]. In figure 16, AP indexes 1-79 depict APs in the class rooms, AP indexes 80-99 depict APs in the administration offices, AP indexes 100-119 depict AP in the library, and AP indexes 120-130 depict APs in the laboratories. Figure 16 shows the fraction of abrupt ending sessions to the total number of sessions for each AP. From results in figure 16, we observe that abrupt ending of AP connections tend to happen in most locations where APs are highly utilized e.g. library and class rooms (see peaks around AP indexes 80 and 100). Additionally, locations that are covered by Cisco C1250 AP appear to have significant impact on the abrupt ending of connections occurrences, see peaks between AP indexes 40-60.

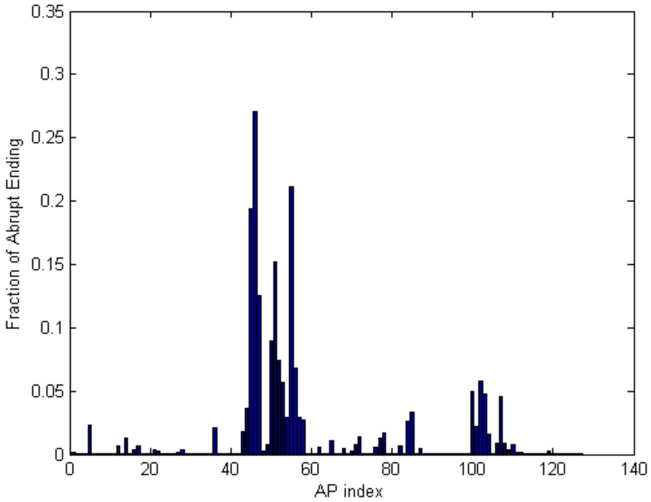
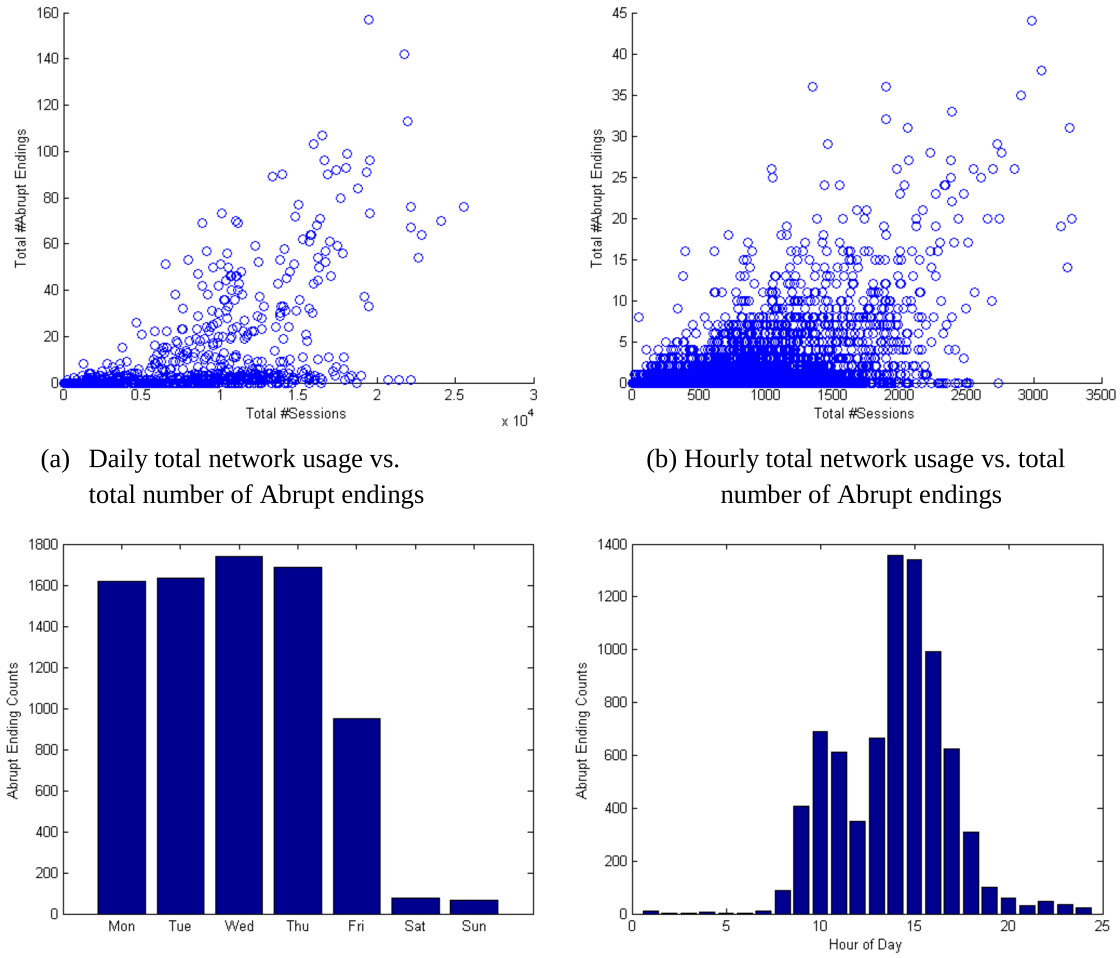


Figure 16. Fraction of abrupt ending sessions to the total number of sessions for each AP

5.3.5 Abrupt Ending and Usage

We lastly aim to understand the impact of total aggregate network usage on abrupt ending of AP connections. From results in figure 17 (a) and (b), abrupt ending of 802.11 AP connections have shown to have a positive correlation to the overall aggregate network usage: with strong daily, hourly, hour of the day and day of the week usages. Their corresponding Pearson correlation coefficients are 0.89, 0.94, 0.96, and 0.93 respectively. Since our data is from a campus 802.11 network with no residential buildings, abrupt endings of connection appear to follow student's usage pattern behavior, with higher counts observed on week-days than on week-ends (see figure 17(c)). Higher counts of abrupt endings are observed particularly in the hours where students are in campus and few in the night (figure 17(d)). These results indicate the higher the aggregate 802.11 network infrastructure is utilized, the higher the likelihood of the abrupt ending of connections to its APs.



(a) Daily total network usage vs. total number of Abrupt endings

(b) Hourly total network usage vs. total number of Abrupt endings

(c) Total number of abrupt endings for each day of the week

(d) Total number of abrupt endings for each hour of the day

Figure 17. Underlying salient features of the aggregate 802.11 total network usage and total abrupt ending occurrences

5.4 Statistical Models of Abrupt Ending Occurrences

5.4.1 Methodology

In this section, we aim first to find out if there exists any significant statistical relationship between abrupt ending occurrences and aggregate 802.11 network usage, in terms of the total number of sessions established per hour. Second, we aim to propose different statistical models that best capture the underlying relationship between abrupt endings and aggregate 802.11 network usages. In this way, models for characterizing abrupt ending of AP connections applicable for fault management of large-scale 802.11 networks can be established.

To accomplish this investigation we first obtain average values of abrupt endings for different ranges of session intervals called bins. We search for abrupt ending event in every 10 session length while maintaining a count. A session interval is considered to be a bin if at least 50 non-zero elements (abrupt ending events) have been identified. This allows us to reduce the randomness of the response variable observed in figure

17(b), while introducing linear effects in the observations. This approach is similar to the partitioning method used in [185]. The resulting plot depicting the relationship between these session intervals (bins) and their respective average number of abrupt endings is given by figure 18. Assuming this linear behavior, traditional linear regression techniques using standard least-squares estimation can be employed.

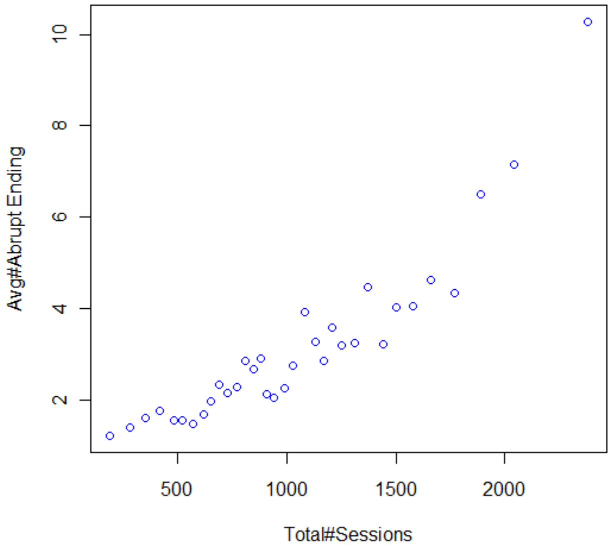


Figure 18. Total 802.11 session count vs. average number of abrupt endings

5.4.2 Linear Regression Model

In this section, we intend to fit a simple linear regression model in order to depict details of the statistical relationship between average abrupt endings *(y)* and the aggregate 802.11 network usage in terms of total number of sessions per hour *(x)*, illustrated by figure 19. We fist perform correlation analysis between the two variables. Pearson correlation analysis of the variables resulted into 0.94, which indicates strong positive correlation between independent variable *"x"* and dependent variable "*y*". We then proceed in fitting linear regression model, traditionally given by the equation: $y = \beta_o + \beta_1 x + \epsilon$. We assume the error term to be independent of *"x"* and is normally distributed variable with zero mean, similar to previous work [186]. Moreover, in our linear model the y-intercept β_o is insignificant, because its p-value (0.486) is much greater than all significant levels tested (0.10, 0.05, 0.01, and 0.001), hence fail to pass goodness of fit test. Interestingly, this result affirms further our first intuition that no abrupt ending can be observed once there are no sessions in the 802.11 networks. We therefore consider $\beta_o = 0$, which simplifies the final linear regression equation to $y = \beta_1 x$.

We use linear model package in R to obtain the estimates for $\beta_1 = 0.0030685$ and its corresponding p-value < 2e-16; which passes goodness of fit test at all significance levels tested (0.10, 0.05, 0.01, and 0.001). The final linear equation for estimating average number of abrupt endings given total number of sessions is given by the following linear equation: $y = 0.0030685\ x$, portrayed visually by figure 19. In order to assess goodness of fit of our proposed linear model we use the coefficient of determination R^2 [187]. The coefficient R^2 indicates how closely values obtained after fitting a model match the dependent variable the model is intended to predict. R^2 values are usually between 0 and 1; a higher value of R^2 implies higher prediction capability of the model. We obtain the estimate for $R^2 = 0.96$, which further confirms the significance of the statistical relationship between these two variables. In simple terms, this means our proposed simple linear regression model is able to approximate 96% of the response in the dependent variable.

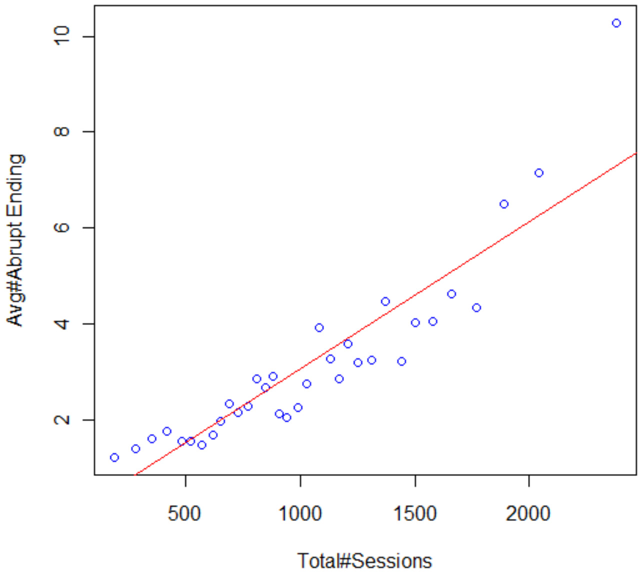
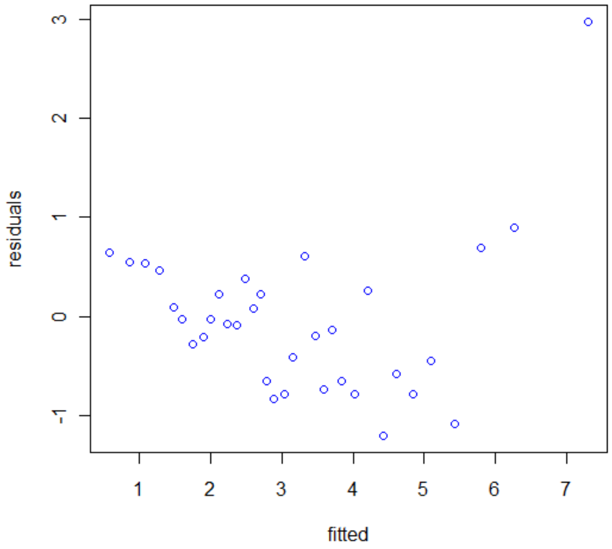
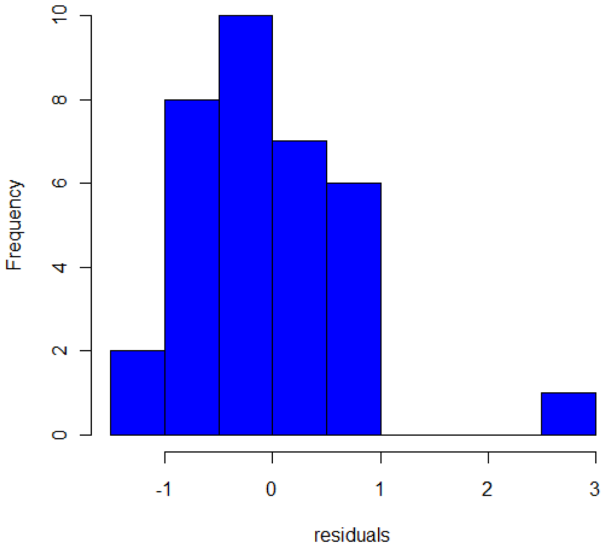


Figure 19. Plot of the proposed simple linear regression model fittings

We lastly examine the residuals and the fitted values of our simple linear regression model. This is a standard way in any modeling endeavor, for gaining further insights related to the goodness of fit of the proposed model. By definition the residual data of the simple linear regression model is the difference between the observed data of the dependent variable and the fitted value. Analysis of fit (figure 20(a) and 20(b)) suggests that the proposed linear equation (figure 19) is able to approximate bulk of abrupt ending averages (about 96%) as majority of the residuals (observed - fitted values) lies between -1 and 1.



(a) Plot of the fitted value vs. residuals after regression line fitting



(b) Histogram of the residuals after regression line fitting

Figure 20. Simple linear regression model fittings and resulting analysis of fit

5.4.3 Continuous Probability Distributions Models

The good fitting results achieved by the regression analysis of the variables in the previous section 5.4.2 ignore the effect associated to distributions of elements within the bins and the possible impact of the error term in the relationship. In this section, we argue that better or improved fitting results can be achieved by employing other statistical models that can as well fit elements within the bins i.e., fitting individual abrupt ending occurrences at distinct intervals of the total number of sessions. A large number of elements within most session intervals lie along the x-axis close to zero (see figure 17(b)). As such, we believe the exponential family of continuous probability distributions models (**Exponential, Gamma, and Gaussian**) is better suited to explain this underlying characteristic. To assess goodness of fit for these models with respect to each bin, we use the Log-likelihood and the Akaike Information Criterion (AIC) metrics similar to chapter 4. These are standard figure of merit in statistical learning [47, 182, 183]. The model with larger log-likelihood value and AIC value is better than the one with smaller log-likelihood value and AIC value for the same set of data.

The log-likelihood of a model's probabilistic density function M with parameter Θ on a data set $X = (x_1, x_2, \dots, x_N)$ is defined as:

$$LL(M; \Theta; X) = \sum_{i=1}^{N} \ln M(x_i; \Theta)$$

Whereas the general form for calculating Akaike Information Criterion (AIC) for a model with log-likelihood (LL) and the number of parameters K is defined as:

$$AIC = -2 * LL + 2 * K$$

Exponential distributions: To commence this modeling effort, similar to chapter 4, we first pick a simple Exponential distribution model. We fit this model on abrupt endings data at each intervals (bins) of the total number of sessions, using maximum likelihood estimation function in R. The following are the PDF and log-likelihood expressions on sample set $X = (x_1, x_2, \dots, x_N)$:

$$p_{exp}(x;\ \lambda) = \lambda \exp(-\lambda x)$$

$$LL = N \ln \lambda - \lambda \sum_{i=1}^{N} x_i$$

Gamma distributions: Similarly in another attempt to gain better fitting, we fit a two parameters model of continuous probability distributions i.e. Gamma distributions on abrupt ending data at each bin. The PDF and corresponding log-likelihood expressions for Gamma distributions on sample set $X = (x_1, x_2, \dots, x_N)$ are:

$$p_{gam}(x; \alpha, \beta) = \beta^{\alpha}/\Gamma(\alpha).\, x^{\alpha-1} \exp(-\beta x)$$

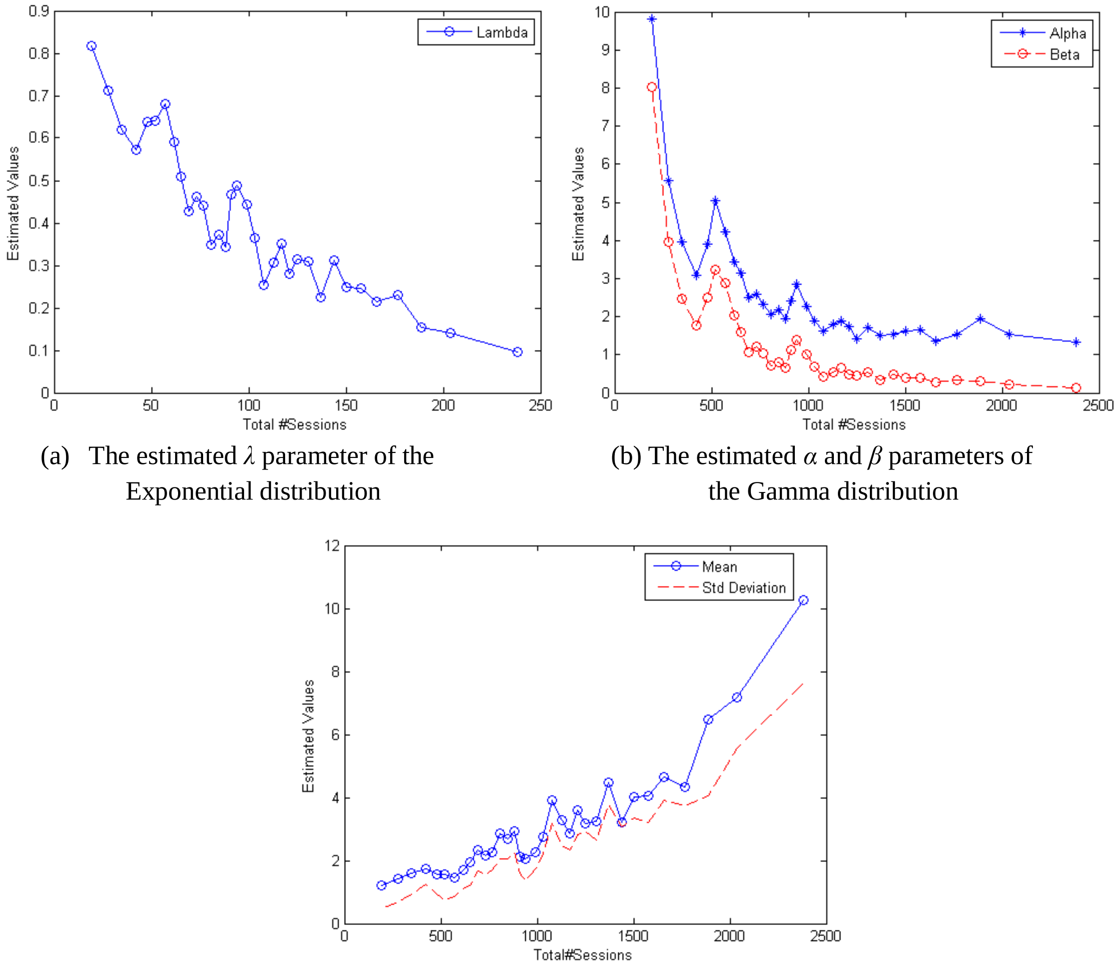
$$LL = \frac{N\alpha}{\beta} - \sum_{i=1}^{N} x_i$$

Gaussian distributions: In a final attempt to gain an improved fitting, we eventually fit Gaussian distributions on abrupt ending data at each bin using maximum likelihood estimation function in R. The following are the PDF and log-likelihood expressions for the sample set $X = (x_1, x_2, \dots, x_N)$ which is normally distributed $N(\mu, \sigma^2)$:

$$p_{gau}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \left(\exp\left(-\frac{(1-\mu)^2}{2\sigma^2} \right) \right)$$

$$LL = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

Results: Figure 21 depicts fitting results of the Exponential, Gamma, and Gaussian distributions at various bins. Figure 21(a) shows the estimated rate (λ) parameter of the Exponential distributions at each bin. When we take the inverse of the estimated rates (i.e. 1/ λ), this yields very similar values to bin's averages used in the linear regression model fitting (figure 19). Figure 21(b) shows the estimated shape parameter α and rate parameter β of the gamma distributions, as estimated by maximum likelihood function in R. Again these parameters when taken as α / β at each bin produce similar values as those used in linear regression fittings. Figure 21(c) shows the estimates for mean μ and standard deviation σ of the Gaussian distributions. Again, the estimates for means μ are very similar in values to the bin's averages used in the linear regression model fitting of as well as $1/\lambda$ of the exponential model and α / β of the Gamma distributions.



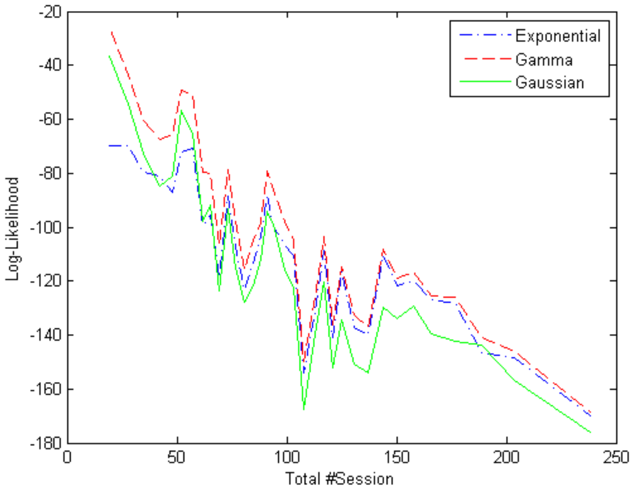
(a) The estimated λ parameter of the Exponential distribution

(b) The estimated α and β parameters of the Gamma distribution

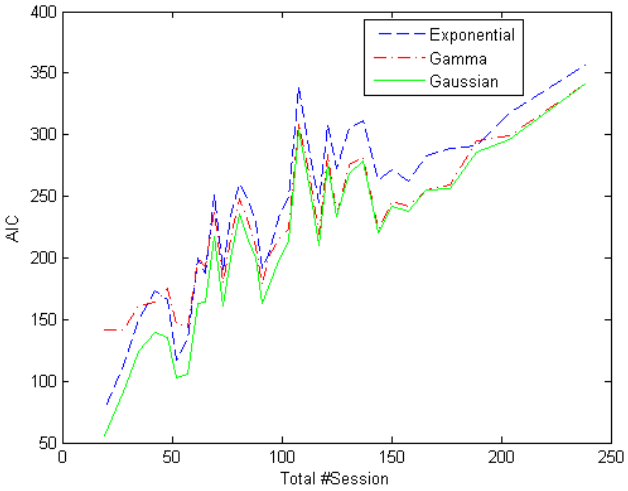
(c) The estimated μ and σ parameters of the Gaussian distribution

Figure 21. Model parameters and fitting results for Exponential, Gamma, and Gaussian distributions models on abrupt ending data set at each bin

In order to determine goodness of fit of the proposed models, we subsequently compute the log-likelihood and AIC values at each (bin) using these estimated rate parameters. Results are presented in figure 22(a) and 22(b), respectively. Generally, Exponential distributions model outperforms Gamma and Gaussian distribution models in terms of AIC values in most of the bins (see dash-dot line in figure 22(b)), while achieving comparable results in terms of log-likelihoods (see figure 22(a)). This is due to the simplicity of the exponential model which has only one parameter. Although Gaussian and Gamma models have similar complexity, in terms of number of parameters, the poor performance of the Gaussian model in terms of log-likelihoods is attributed to skewedness of elements within bins (i.e. elements being not symmetric); this underlying characteristic can be better explained by Exponential and/or Gamma model.



(a) Log-likelihood results for all the three models



(b) AIC results for all the three models

Figure 22. Log-likelihood and AIC results for Exponential, Gamma, and Gaussian distributions models on abrupt ending data set at each bin

5.5 Conclusions

In this chapter, we have identified a new usage pattern named “abrupt ending” of 802.11 AP connections that happens when a large number of user’s sessions in the same access point (AP) end within a one second window. We observed up to 40 distinct user’s sessions ending at the same second in some access points and over thousands of abrupt endings in a two and a half year 802.11 traces of the Faculty of Engineering of the University of Porto. We confirmed the existence of significant statistical relationship between abrupt ending occurrences and 802.11 aggregate usage, in terms of total number of sessions. We proved this relationship by: 1) Pearson correlation coefficient between variables (r = 0.94), 2) Coefficient of determination (R^2 = 0.96) between variables in the linear regression model, and 3) P-values (< 2e-16) which passes all significant levels tested. In addition, we demonstrated that a family of continuous probability distributions (e.g. Exponential, Gamma, and Gaussian models) is suitable to capture the underlying relationship between abrupt ending occurrences and aggregate 802.11 usages. Due to its simplicity, Exponential distributions model outperformed others and proved to be sufficient for capturing the overall characteristics between these variables.

We believe our methods and models presented in this chapter can enhance anomaly detection in the large-scale 802.11 networks. Detecting patterns of anomalies in AP usage is essential for effective management of large-scale networks, particularly in handling network reliability problems. Based on the insights of the models proposed in this chapter, 802.11 administrators might be able to characterize abrupt ending occurrences in large-scale 802.11 infrastructures and thereby plan, and make decision or act in an attempt to guarantee continuous and reliable coverage of 802.11 networks. In the next chapter, we will present algorithm for characterizing abrupt ending of AP connections discussed in this chapter into different forms of anomaly-related patterns. Once the appropriate nature of an anomaly is established, relevant counter measures can be easily taken to remedy such anomaly. This is important for a proactive approach to fault management of the large-scale 802.11 networks.

Chapter 6

Detecting and Modeling Patterns of Abrupt Ending of 802.11 AP Connections

6.1 Introduction

The task of anomaly detection plays significant role in 802.11 networks. In most cases, the detection of patterns of anomaly results in actionable decision by network administrators in an attempt to mitigate the effects caused by anomalous situations. For network administrators to make informed and appropriate decisions first the true nature of anomalies needs to be established and existence of their underlying patterns identified. In this chapter, we take a step further from detecting individually abrupt ending of AP connections and focus on detecting and characterizing patterns of anomaly resulting from the occurrence of AP abrupt ending of connections. We investigate the existence of these patterns by analyzing the timing of each abrupt ending event, the regularity of AP abrupt ending within a day, and the presence and absence of continued sessions after each AP abrupt ending event. We consider these factors mainly for detecting AP-related anomalous patterns namely interference across AP vicinity, AP persistent interference, AP halt/crash, AP overload, and AP interference. For detecting the user-related patterns of user's authentication failure and users' intermittent connectivity to APs, we analyze the identity of users involved in each abrupt ending, and also we observe user intermittent sessions just before abrupt ending occurrences and their prevalence afterwards.

The emphasis on detecting and characterizing these anomaly-related patterns is crucial for fault management of the large-scale 802.11 networks. Proper detection of the aforementioned patterns might help network administrators in making informed decisions and taking timely actions. Examples of the short term actions that can be taken include rebooting of the APs, adjusting power settings of APs, and changing of operational channels and antenna radiation patterns of the APs [129-131]. When thinking of long term planning and maintenance of 802.11 networks, actions such as adding more APs, and replacing and relocating some APs can be useful given the nature of the problem. In this chapter, we do not further explore actionable decisions by 802.11 administrators, but rather focus on characterizing abrupt endings into respective patterns of AP usage anomaly.

In section 6.2, we present an algorithm for detection and characterization of abrupt endings into different forms of anomaly-related patterns. In section 6.3, we discuss experimental results using different thresholds of time intervals learned from our data set, for depicting anomaly-related patterns described in section 6.2. In section 6.4, we propose statistical models for characterizing occurrence of these anomaly-related patterns. In section 6.5, we provide evaluation of our experimental results using a density based clustering algorithm (DBSCAN) and on data set from different 802.11 deployments. In section 6.6, we provide an online implementation of the detection and characterization of abrupt endings into their respective anomaly-related patterns using complex event processing techniques. In section 6.7 we provide concluding remarks.

6.2 Algorithm for Detection and Characterization of Anomaly-related Patterns

6.2.1 Anomaly-Related Pattern Characterization

We consider the following anomaly-related patterns based on manual inspection of 802.11 AP usage data. From AP's perspective these patterns are: 1) Interference across AP vicinity patterns - this is when abrupt ending occurred as well to neighboring APs within specified time interval, possibly due to interference from the neighboring APs or the increase in number of collisions taking place in the wireless medium. 2) AP persistence interference patterns – this is where repeated abrupt endings are observed for a given AP during a specified duration of time, possibly due to dead spot/RF holes. 3) AP overload patterns - if the presence of continued session(s) to an AP is evident after abrupt ending during a specified time interval; this might be the case of heavy utilization by some few users. 4) AP halt/crash patterns - when no log event (START, ALIVE, or STOP) is seen after abrupt ending for a given time period; this may indicate possibility of AP failure. 5) AP Interference patterns - where abrupt ending events do not belong to any of the aforementioned patterns, perhaps caused by the presence of source of interferences e.g. RF devices. We name these anomalous patterns as AP-related patterns.

We also consider user-related anomalous patterns in our study. These patterns are: 6) User authentication failure to AP – when abrupt ending resulted from a single user, possibly due to the use of wrong or expired credentials by user or problem in the network services e.g. authentication serves, and 7) User intermittent connectivity to AP patterns – when users intermittent connectivity is evident just before and after abrupt ending occurrence, possibly due to changing wireless conditions and inconsistent coverage of 802.11 networks.

6.2.2 Anomaly-Related Pattern Definition

We use the following definitions for detecting anomaly-related patterns outlined in the previous subsection. For AP-related patterns: 1) Pattern A (interference across AP vicinity) occurs when more than one AP encounter abrupt endings in a 60 seconds time interval. 2) AP Persistent interference (pattern B) happens when four or more abrupt endings occur in one specific AP in a space of one day. 3) AP overload (pattern C) happen when continued sessions are observed within 15 minutes after abrupt ending. 4) AP halt/crash (pattern D) occurs when an AP stops registering events for more than half an hour after AP abrupt ending occurrences, i.e. when no user connection is observed at an AP within half an hour after abrupt ending. 5) AP interference (pattern E) happens if abrupt endings do not belong to any of the 1-4 aforementioned patterns.

For user-related patterns: 6) Pattern F (user authentication failure) occurs when abrupt endings emerged from a single user for a given AP. 7) Pattern G (User intermittent connectivity) occurs when unstable user connections are observed before and after abrupt endings. Unstable user connections are sessions with short duration i.e. <= 120 sec. For all these patterns we chose values (thresholds) based on the manual observation of the data set, as will be illustrated in section 6.3.

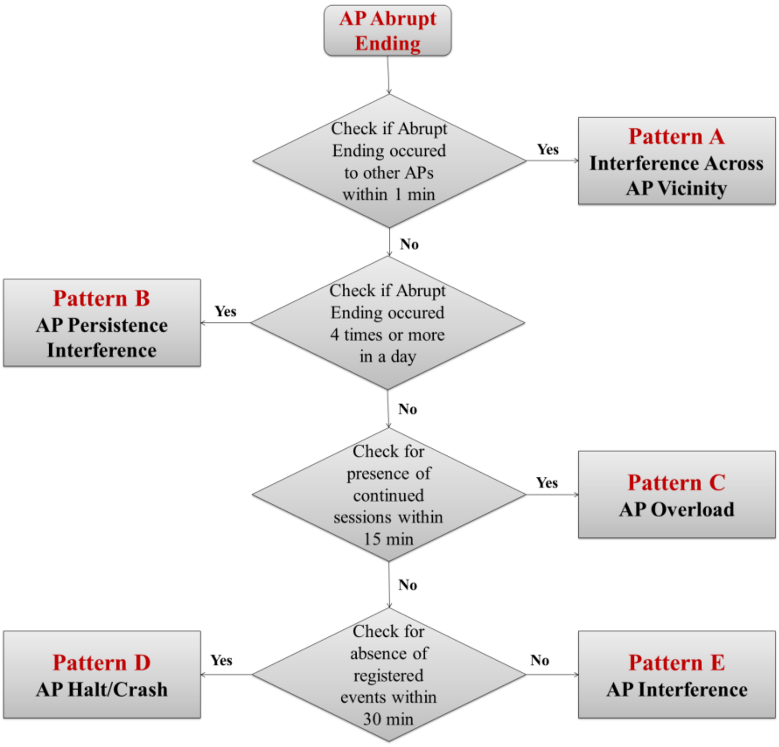

AP Abrupt Ending
Check if Abrupt Ending occured to other APs within 1 min
Yes
Pattern A
Interference Across AP Vicinity
No
Check if Abrupt Ending occured 4 times or more in a day
Yes
Pattern B
AP Persistence Interference
No
Check for presence of continued sessions within 15 min
Yes
Pattern C
AP Overload
No
Check for absence of registered events within 30 min
Yes
Pattern D
AP Halt/Crash
No
Pattern E
AP Interference

Figure 23. Algorithm for detection and characterization of anomaly-related patterns

6.2.3 Detection Algorithm Overview

Figure 23 depicts the algorithm we propose for offline detection and characterization of anomaly-related patterns of AP usage. When an abrupt ending is observed at timestamp t_1, we check if abrupt endings also happened to other APs within 60 seconds of the first abrupt ending. If this is the case, then we claim pattern A (interferences across AP vicinity). If only one AP encountered abrupt ending in an interval of 60 seconds, then we check if this AP had encountered other abrupt endings during a day (24 hours period). If it happens more than four times, then we claim pattern B (AP persistent interference). Otherwise, we check AP for the pattern C (AP overload). If we observe continuing session(s) within 15 minutes of the abrupt ending occurrence, then we declare pattern C. If pattern C is not declared, we check the possibility of AP crash/halt pattern D, i.e. if no single event is observed for a period of half an hour after the abrupt ending. When none of the aforementioned conditions are met then pattern E is declared (AP interference).

Our proposed algorithm is capable of effectively characterizing AP-related patterns (A-E) once supplied with the appropriate input data (i.e. all detected AP abrupt ending events on the data set); on the other hand, our algorithm is not well suited for depicting user-related patterns (F and G), given the same input data. The detection of these patterns requires additional user information to be supplied alongside each abrupt ending event, and its inclusion may potentially increase the complexity of the algorithm. For example, pattern F detection requires identity of users (User ID) involved in the abrupt ending so as to determine if the abrupt ending had emerged from a single user. Whereas pattern G detection demand accessibility of distinct user intermittent sessions just before and after AP abrupt ending. In future work we consider incorporating these patterns in our algorithm.

The main advantage of using our algorithm is that 802.11 network administrators can easily detect the presence of anomalies in the usage of APs in an 802.11 infrastructure, and most importantly their respective true nature (e.g. AP overload, AP halt/crash, and AP interference). Consequently, relevant effective counter measures can be considered depending on the nature of anomaly, in an attempt to assure reliable coverage of large-scale 802.11 networks.

6.3 Anomaly-Related Pattern Detection: Experimental Results

6.3.1 Interference Across AP Vicinity Patterns

For detecting interference across AP vicinity patterns we tested different thresholds of time intervals, as shown in figure 24. We define AP vicinity segment to be the set of abrupt endings for APs that lie within 60 seconds of each other. The total number of vicinity segments detected and the total number of interfering instances (APs) within each segment for each threshold are depicted in figure 24. An instance of one AP vicinity segment may comprise APs ranging from a minimum of 2 APs. The maximum number of APs we observe in one vicinity segment is 30 APs.

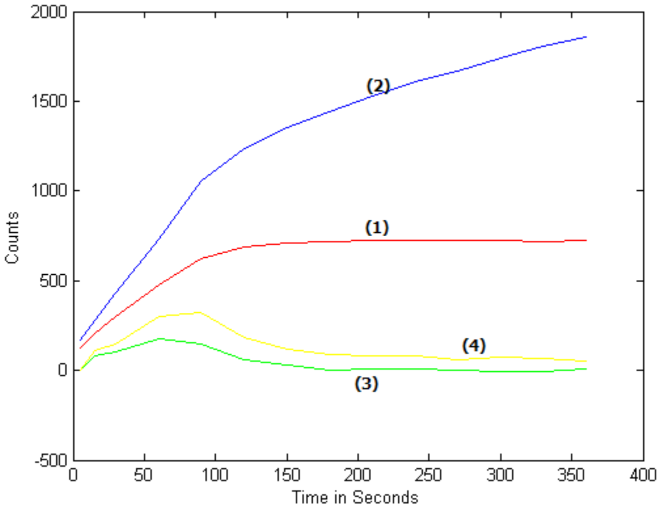


Figure 24. Different threshold settings tested vs. counts: (1) total number of vicinity segments detected, (2) total number of interfering APs within vicinity segments, (3) difference in number of segments from previous time threshold, and (4) difference in number of interfering APs between consecutive segments

Figure 24 shows that as the threshold increase the total number of AP vicinity segments and total number of interfering APs within segments increases, although after 60 seconds the number of segments increases more slowly. However, increasing the threshold interval could lead to involving APs in vicinity segments which in fact are not neighbors, on the other hand smaller threshold may lead to splitting of one AP vicinity segment into more than one distinct segment. We choose the threshold of 60 seconds since the difference in vicinity segments and interfering instances at consecutive intervals peaked at 60 seconds and insignificant increase in the number of vicinity segments detected is observed afterwards (see line #1 in figure 24). Adequate threshold settings here can assist network administrators in identifying APs that tend to interfere with each other in the 802.11 infrastructure.

6.3.2 AP Persistent Intereference Patterns

For detecting these patterns we count the number of AP persistent interference instances detected while varying detection thresholds, as indicated in table 14. Using this technique we aim to determine a reasonable threshold for depicting cases signifying severity of abrupt ending occurrences to individual APs in the 802.11 infrastructure.

#Abrupt endings per AP per day	#AP Persistence Interference Patterns	Reduction Ration
>= 2	1711	-
>= 3	854	2.004
>= 4	469	1.821
>= 5	258	1.818
>= 6	143	1.804
>= 7	80	1.787

Table 14. Thresholds tested for detecting AP persistence interference

Table 14 shows that as the number of abrupt endings per AP per day increases, the number of instances of persistent interference detected appear to decrease, with small change of the number, i.e. previously detected instances – currently detected instances, noted beyond the frequency of 4. Because of this, number of abrupt ending per AP per day >= 4 was chosen as appropriate threshold for detecting AP persistence of interference. Another argument supporting this threshold is the reduction ratio in table 14, (i.e. ($\frac{n+1}{n}$), n = #Abrupt endings per AP per day) which does not change much after frequency of 4. In our data set, AP persistence of interference ranges from the minimum of 4 abrupt endings per AP in one day to the highest frequency observed of 18 abrupt endings in one day for one particular AP. Appropriate threshold settings here can help to identify APs which are prone to interference in the large-scale 802.11 networks.

6.3.3 AP Overload Patterns

For detecting AP overload patterns we check for the presence of continued session(s) at different time intervals after abrupt endings, as indicated in table 15. In this way, the existence of continuing session(s) after AP abrupt ending can be clearly observed. Detection results for each threshold tested are shown in table 15.

Time interval after Abrupt ending (Mins)	**#AP Overload Patterns**
5	168
10	198
15	204

Table 15. Thresholds tested for detecting AP overload patterns

Table 15 shows that the number of AP overload patterns increases as the time interval after AP abrupt endings increase, and reaches maximum at 15 minutes. To remove any uncertainty in detection, we choose the maximum i.e. 15 minutes as the threshold for depicting any possibility of the presence of continuing sessions after AP abrupt ending. The choice for this threshold takes into account the fact that *"ALIVE"* log events are normally generated every 15 minutes for refreshing user connections to AP (according to the configuration of our APs); thus any user session which started just before occurrence of AP abrupt ending, if that session continued, its ALIVE event can also be captured by this threshold. Otherwise it would have been impossible to capture these sessions. Adequate threshold settings here can help to identify overloaded APs within the large-scale 802.11 network.

6.3.4 AP Halt/Crash Patterns

For detecting AP halt/crash patterns we test different thresholds of time intervals as indicated in table 16. We check for APs that did not register any log event ("START" or "ALIVE" or "STOP") after abrupt ending occurrence within these specified time periods. Our aim is to understand which time interval is appropriate to declare if AP is halted or crashed, due to the absence of registered events after AP abrupt ending.

Time interval after Abrupt ending (Mins)	**#AP Halt/Crash Patterns**
15	33
30	19
60	17

Table 16. Thresholds tested for detecting AP halt/crash patterns

From table 16, we notice that the number of AP halt/crash instances decreases with the increase in time interval and the difference, i.e. previously detected instances – currently detected instances, is insignificant between 30 minutes and 60 minutes. This insight leads to the choice of 30 minutes after abrupt endings to be considered as an adequate threshold for detecting the possibility of AP halt or crash. Adequate settings of threshold here can help to identify crashed or halted APs in the large-scale 802.11 infrastructure networks, hence minimizing their potential impact. Typically users in the vicinity of halted/crashed AP experience frequent disconnections of their sessions, mostly due to inconsistent coverage provided by nearby APs with respect to their positions.

6.3.5 AP Interferences Patterns

We detect interference patterns after removing abrupt endings related to 1) AP persistence interference patterns, 2) across AP vicinity patterns, 3) AP overload patterns, and 4) AP halt/crash patterns. In this case, abrupt endings that do not belong to any of the aforementioned patterns were considered a default case of AP interference patterns. This is in accordance to our proposed algorithm (figure 23), since an abrupt ending by itself is an indicator of AP usage anomaly. From our data set we observe 5536 cases of AP interference patterns. Table 17 depicts frequency of occurrences of abrupt endings of these patterns to individual APs per day.

#Abrupt endings per AP per day	#Interferences Patterns
1	3803
2	1042
3	691

Table 17. AP Interference patterns occurrence

Results from table 17 shows that more than half (i.e. 61%) of abrupt endings related to AP interference patterns occurred only once to individual APs in a day, while 26% occurred twice in a day, and 13% happened three times in a day. Remember, any abrupt endings per AP per day beyond the frequency of four would be considered as AP persistent interference patterns (pattern B), hence excluded from AP interference patterns (pattern E). Detection of AP interference patterns might help to identify locations in 802.11 infrastructure with the possibility of having temporary sources of interference close to the vicinity of APs: for example microwave oven, motion sensors, industrial and medical RF devices etc.

6.3.6 User Authentication Failure Patterns

We detect user authentication failure patterns by observing identity of users involved in the abrupt endings. This is so in order to depict cases of abrupt endings that resulted from exact one single user i.e. a user who have three or more sessions stopped at exactly the same second in the same AP. In most cases this could only be possible if a user is unable to associate to 802.11 networks for some reasons.

From our data set we observe 59 cases of authentication failure patterns involving 41 APs, where users start associating to one AP and being disconnected at almost the same second. In a normal 802.11 operational environment, these patterns can actually hint at invalid or expired credentials supplied by users and/or perhaps problem with the logging severs or network services. Detection of these patterns is crucial in order to deal with problems facing individual users of the large-scale 802.11 networks, and also for troubleshooting problems pertaining to network services in the wired side of the 802.11 network.

6.3.7 User Intermittent Connectivity Patterns

To detect these patterns we observe user's intermittent connections (i.e. frequency of user sessions disconnection to 802.11 AP) just before abrupt ending occurrence and subsequently after. In this context, user's intermittent connectivity to AP or user's frequent sessions disconnection refers to very short user sessions of typically less or equal to 120 seconds, i.e. we look for STOP - START <= 120 seconds. We use 120 seconds because most of our campus 802.11 users are laptop users with small degree of mobility. It is rare in a normal working condition these laptop users would want to connect and use 802.11 networks for 120 seconds or less. User's intermittent connectivity to APs can be the result of RF interferences, inconsistent coverage, or changing wireless conditions.

To depict these patterns we observe stability of user's connections five minutes before the abrupt ending occurrence and five minutes after. Our aim is to understand if there were any indications of connectivity problems well before abrupt ending occurrence, and importantly their perseverance afterwards. Depending on the frequency of user sessions disconnection to AP, we classify intermittent user session disconnections into four broad categories: 1) none - if no user frequent disconnection is observed, 2) low - if users are frequently disconnected between one and 5 times, with an average of about one frequent disconnect per minute, 3) high - if users are frequently disconnected between 6 times and 10 times, with an average of about two frequent disconnect per minute, and 4) very high - when users are frequently disconnected more than 10 times, with an average of three disconnect per minute or more.

Frequent User Session Disconnections	**Total # Before AE**	**Total # After AE**	**Percentage Before AE**	**Percentage After AE**
None	5048	4565	65	63
Low	149	205	4	4
High	425	614	8	9
Very High	1652	1890	23	24

Table 18. Characterization of intermittent connectivity patterns before and after abrupt endings

Table 18 shows that about 65% of abrupt endings did not show any indication of user's connectivity problems well before abrupt ending, while 35% showed indications of user's connectivity problems before abrupt ending occurrence. 63% indicates there was no users' connectivity problem after abrupt ending, while 37% indicates prevalence of disconnections after abrupt endings occurrence. These statistics can help to understand the general state of a large-scale 802.11 network in terms of stability of user's connections to APs, mostly before and after abrupt endings occurrences.

6.3.8 Summary

In this section, we presented experimental results for all anomaly-related patterns investigated in this thesis. We demonstrated the impact of each threshold setting on the detection results for each pattern. Detection of these patterns may necessitate remedial decision and action from 802.11 administrators. Although actionable decisions are beyond the scope of this thesis, but here we highlight some of the remedial actions that can be considered for each of these patterns. For example: 1) pattern A - may include action such as adjusting power settings of APs, and changing of radiation patterns and frequency channels of the APs. 2) Pattern B - may require relocating APs that often encounters abrupt ending, or changing configuration parameters of these APs. 3) Pattern C - may need action such as increasing density of APs, especially in the location(s) which appears to be congested so often. 5) Patten D - may require rebooting, or replacing the crashed or halted APs. 5) Pattern E - in addition to removing source of interference close to APs, requires actions similar to pattern A; since both are caused by RF interference. 6) Pattern F - may necessitate blocking of unauthorized or problematic users, issuing of new credentials to legitimate users, and in some situations carefully examination of the wired part of the large-scale 802.11 networks can be desired. 7) Pattern G - decisions here depends on the nature of the abrupt ending characterizing the identified intermittent connections: for example, if the abrupt ending belongs to pattern A then relevant measures applicable to pattern A can be employed, and so on. These remedial actions are attempts that can be considered by 802.11 administrators, in order to guarantee continuous and reliable coverage of the large-scale 802.11 networks.

6.4 Anomaly-Related Pattern Modeling: Experimental Results

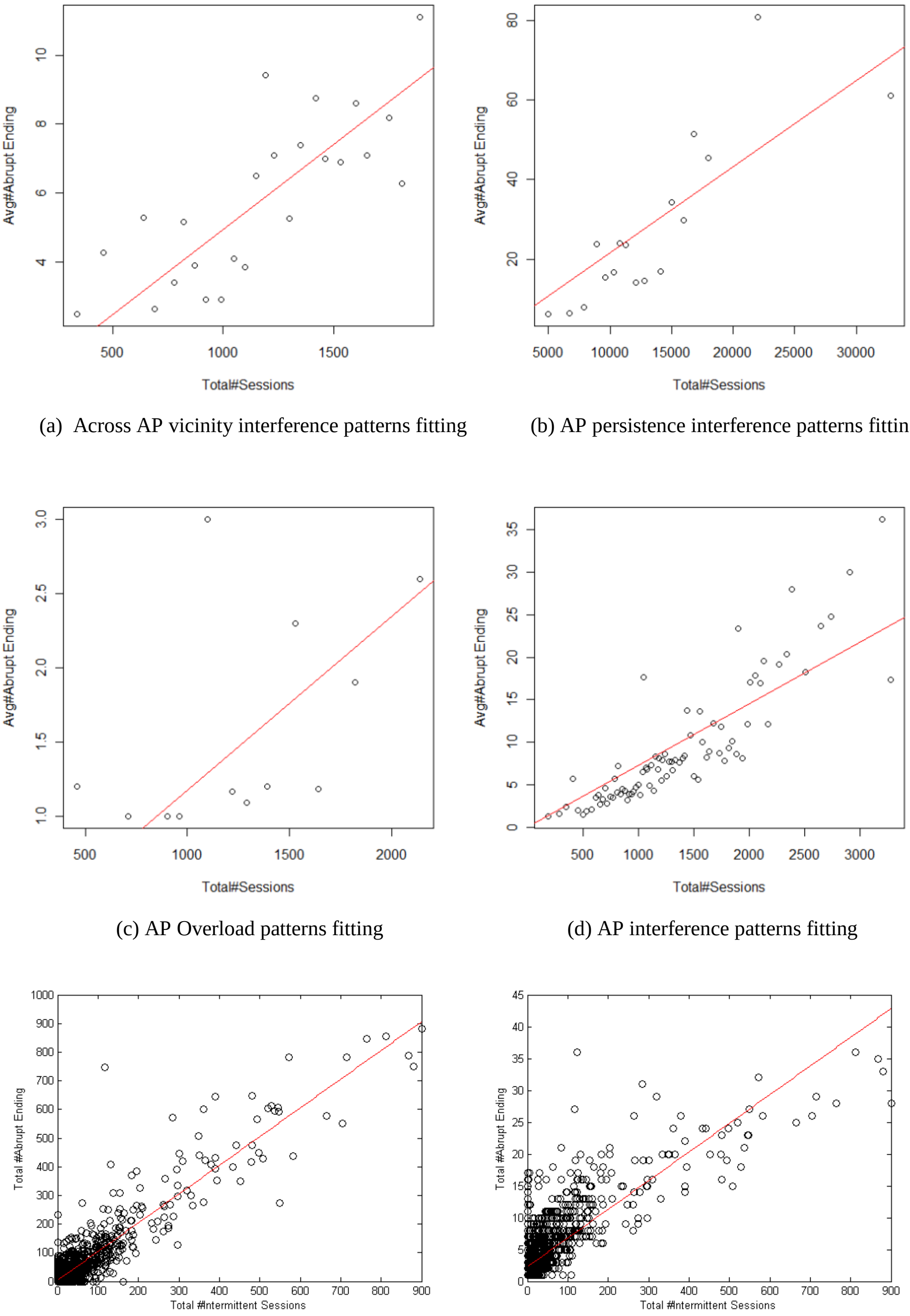
6.4.1 Linear Regression Models

In this section, we propose simple linear models to capture overall characteristics of the abrupt endings associated to anomaly-related patterns investigated in section 6.3 and aggregate 802.11 network usage, in terms of the total number of sessions established per hour, similar to section 5.4.2. Although in section 5.4.2 we used average abrupt endings as a dependent variable irrespective of patterns, in this section we use average abrupt endings significant to a particular pattern as dependent variable *"y"* and aggregate network usage (i.e. total number of sessions per hour binned at various session intervals) as independent variable *"x"*. However, there is an exception here particularly for anomaly-related pattern B (AP persistent interference), because these patterns are detected in a period of one day (see definition in subsection 6.2.2). Rather than using hourly aggregate network usage for these patterns, we instead use daily aggregate 802.11 network usage (total number of sessions per day) as independent variable *"x"*. Another exception is for pattern G (user intermittent connectivity) where we analyzed un-binned observations of the total number of intermittent sessions (see definition in subsection 6.2.2) per hour before and after abrupt ending occurrences (as independent variable *"x"*) versus the total number of abrupt endings in the same hour (dependent variable *"y"*). We did not bin observations for this pattern because the majority of elements in the response variable are non-zero with linear behavior (see figure 25(e) and 25(f)). We fit linear regression model between variables for each anomaly-related pattern investigated in this chapter using equation $y = \beta_o + \beta_1 x$ similar to chapter 5. Table 19 present parameters of the fitted simple linear regression models and their corresponding goodness of fit tests for all anomaly-related patterns of AP usage. Figure 25 illustrates simple linear regression models fitting only for those patterns exhibiting significant relationship between variables.

Results: Pearson correlation (*r*) and regression analysis of the variables (R^2 and p-values in table 19) indicates strong significant statistical relationship between variables for all anomaly-related patterns studied except for two patterns: 1) AP halt/crash and 2) user authentication failure. In these two patterns the Pearson correlation and R^2 coefficients are reasonably small, whilst their p-values fail to pass all significant levels tested, particularly at 0.01 and 0.001. These results means AP halt/crash patterns and user authentication failure patterns might be the consequence of other causes apart from 802.11 network usage. For example: AP misconfiguration and defective hardware for the former pattern, while for the later pattern might be the consequence of the use of wrong or expired credentials and of problems related to network or network services in the wired part of 802.11 networks, e.g. authentication/logging-in servers problem. As for the other patterns, results suggest statistical significance of the relationship between variables in the regression equations, since their p-values pass all significant level tested (0.10, 0.05, 0.01, and 0.001). In addition, results for R^2 indicate that all our fitted linear models are able to explain the majority of the response in the observed values of the dependent variable. Note, the y-intercept is insignificant for all anomaly patterns studied, similar to the insights of chapter 5, except for pattern G (intermittent connectivity patterns before and after abrupt ending occurrences), where in both cases the y-intercept is significant with their p-values being able to pass all significant level tested (0.10, 0.05, 0.01, and 0.001). These insights indicate that intermittent sessions can be important indicator of 802.11 AP usage anomalies, and can subsequently lead to abrupt ending of AP connections. Therefore, models developed based on these patterns can be helpful in estimating occurrences of AP abrupt ending of connections with respect to usage of these large-scale 802.11 infrastructures.

	Pearson Correlation	**Linear Regression Model Parameters**			
Anomaly-Related Pattern	***r***	**β_0**	**β_1**	**R^2**	**p-value**
Across AP Vicinity Interference	0.87	-	0.0494	0.944	6.68E-13
AP Persistence Interference	0.84	-	0.0216	0.888	5.19E-09
AP Overload	0.74	-	0.00117	0.863	4.38E-03
AP Halt/Crash	0.34	-	0.00219	0.509	0.0358
AP Interference	0.87	-	0.00725	0.898	2.2E-16
User Authentication Failure	0.21	-	0.00083	0.524	0.0264
Intermittent Connectivity (Before)	0.81	2.321	0.0447	0.761	2.2E-16
Intermittent Connectivity (After)	0.84	2.202	0.0432	0.798	2.4E-16

Table 19. Linear regression models fitting results and goodness of fit analysis for each anomaly-related pattern



(a) Across AP vicinity interference patterns fitting

(b) AP persistence interference patterns fitting

(c) AP Overload patterns fitting

(d) AP interference patterns fitting

(e) Intermittent connectivity patterns fitting before abrupt ending

(f) Intermittent connectivity patterns fitting after abrupt ending

Figure 25. Linear Regression models fitting for the significant anomaly-related patterns

6.4.2 Continuous Probability Distributions Models

In this section, we use continuous probability distributions models (namely the Exponential, Gamma, and Gaussian distributions) for capturing overall insights of the abrupt ending occurrences of each anomaly-related pattern given aggregate 802.11 total network usage, similar to section 5.4.3. Fitting results for these models on each anomaly-related pattern are presented in table 20. Results indicate Gamma and Gaussian models achieve slightly better fittings in terms of log-likelihoods over the exponential model, although the small gain attained by these models comes at the expense of the increase in complexity, in terms of number of parameters of the models. This claim is further justified by the AIC results, where exponential model appear to have slightly better results in terms of AIC values for most patterns. Given these very close results (*LL* and *AIC* values) by the virtue of simplicity of the model, simple exponential distribution is favored for modeling all anomaly-related patterns investigated in this chapter.

	Exponential Model			Gamma Model				Gaussian Model			
Anomaly-Related Pattern	**λ**	**LL**	**AIC**	**α**	**Β**	**LL**	**AIC**	**μ**	**σ**	**LL**	**AIC**
Across AP Vicinity Interference	0.171	-66.422	134.843	6.03	1.029	-53.532	111.065	5.857	2.321	-54.271	112.542
AP Persistence Interference	0.036	-73.533	149.066	2.11	0.076	-71.292	146.584	27.812	20.217	-75.233	154.466
AP Overload	0.643	-17.284	36.569	6.268	4.035	-10.635	25.270	1.553	0.679	-12.385	28.770
AP Halt/Crash	0.714	-13.364	28.729	4.442	3.173	-9.304	22.608	1.4	0.8	-11.958	27.916
AP Interference	0.109	-264.121	530.241	2.071	0.225	-253.776	511.553	9.216	7.055	-276.561	557.123
User Authentication Failure	0.932	-18.861	39.723	20.145	18.779	-13.012	29.77	1.4	0.8	-15.747	35.493
Intermittent Connectivity (Before)	0.0182	-611.999	1225.99	0.729	0.0133	-607.453	1218.91	54.821	100.99	-608.96	1221.92
Intermittent Connectivity (After)	0.016	-605.933	1213.87	0.738	0.012	-601.91	1207.82	63.32	110.94	-603.256	1210.51

Table 20. Model Parameters, LL, and AIC values for each statistical model on each anomaly-related pattern

6.4.3 Summary

In this section, we examined the existence of significant statistical relationship between anomaly-related patterns occurrences and 802.11 aggregate usage. We find the relationship to be statistically significant for all anomaly-related patterns examined, except for AP halt/crash patterns and user authentication failure patterns. In addition, we proposed a family of continuous probability distributions e.g. Exponential, Gamma, and Gaussian models to capture underlying relationship between anomaly-related patterns occurrences and aggregate total 802.11 network usage. Despite its simplicity, the Exponential distribution model proved to be sufficient for capturing the overall relationships between variables for all the anomaly-related patterns investigated in this chapter. However, Gamma and Gaussian models achieved very close results to exponential model (see *LL* and *AIC* results in table 20), but complexity in terms of number of parameters counted against them.

6.5 Anomaly-Related Pattern Detection: Experimental Evaluation

6.5.1 Using Data sets from different Hotspots

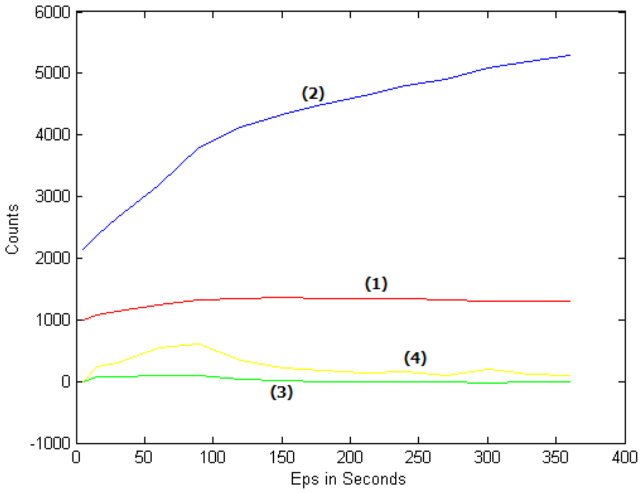
We applied our offline algorithm of section 6.2 (figure 23) using a more recent FEUP trace data set, i.e. 395 consecutive days between 1st May 2010 and 31st May 2011. We suppose the 802.11 infrastructure was more mature in its usage considering the trend stipulated in Table 11. The network was covered by total of 244 APs. During this period we observe more usage on the 802.11 infrastructure (total number of sessions established 4,303,517 i.e. about 46% increase per year), a larger number of users (24,807 i.e. about 37% increase per year), and 14,784 of the abrupt endings i.e. 57% increase per year. Using our proposed algorithm we detected 1201 across AP vicinity interference patterns, 657 patterns of AP persistence interference, 586 patterns of AP overload, and 43 patterns of AP halt/crash. In addition, we detected 452 cases of authentication failure patterns.

We also applied our algorithm to a data set from another university, i.e. University of Minho (UMinho). This data was obtained in the context of FCT project SUM. Our aim is to cross check the existence of abrupt endings and their associated anomaly-related patterns in other 802.11 deployments. We analyzed one month data set from one academic semester (01-30 June 2011). The network had 971 APs and 8,055 users. The total number of sessions established by these users in one month is 1,240,594. We detected 217 AP abrupt ending events, 94 patterns of AP vicinity interferences, 12 patterns of AP persistence interference, 31 patterns of AP overload, 3 patterns of AP halt/crash, and 37 cases of authentication failure patterns.

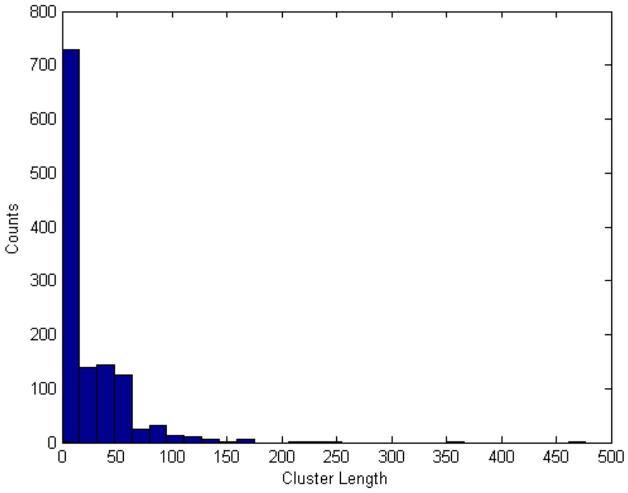
Furthermore, we applied our algorithm to another 802.11 hotspot at INESC TEC with 75 APs, 464 users, and a total number of sessions established in one month of 11,480. From these figures (if compared to the other two 802.11 deployments) we could understand the network is relatively small and not heavily utilized. We did not detect any AP abrupt ending of connections event during the studied one month trace period. This further confirms our intuition about the correlations that exist between abrupt endings and 802.11 network usage, as described in chapter 5: the higher 802.11 infrastructure is utilized in term of network usage, the higher the possibility of abrupt endings and their respective patterns and vice versa.

6.5.2 DBSCAN Clustering Results

We evaluate pattern A (across AP vicinity interference) results of our anomaly detection algorithm of section 6.2, using density based spatial clustering algorithm (DBSCAN) [188]. We take timestamps corresponding to each AP abrupt ending event occurrence as an input to the DBSCAN algorithm. We set minimum number of points in a neighborhood of each point (`MinPts`) to 2, aiming to obtain clusters with the minimum of 2 elements. In this way, we can depict clusters whereby at least two APs were involved in across AP vicinity interference. Consequently, we can establish the minimum number for any vicinity segment detection. We vary the maximum radius of the neighborhood reachability distance (i.e. `Eps`) between elements in the same way as in the detection of across AP vicinity interference patterns (cf. figure 24). The number of clusters and the number of clustered elements for each `Eps`, as well as the difference in number of clusters and clustered elements between consecutive `Eps` are depicted by figure 26.



(a) Different Eps settings for DBSCAN vs. Counts: 1) total number of clusters, 2) total number of elements within clusters, 3) difference in number of clusters, and 4) difference in number of elements within clusters at consecutive intervals



(b) Histogram of the clustered segment lengths

Figure 26. DBSCAN clustering results of the abrupt ending data

Figure 26(a) shows that as the `Eps` in DBSCAN algorithm increases, the number of depicted clusters increase with the significant increase noted at 60 seconds (line #1). Thereafter, the number of depicted clusters did not show any significant increase. These observations match the insights of figure 24, confirming 60 seconds as an optimal threshold for across AP vicinity interference pattern detection. Another argument supporting this threshold is the histogram of the cluster's segment length of the DBSCAN algorithm figure 26(b), which indicates significant change in clusters segment lengths after 60 seconds.

In this section, we presented additional evaluation of pattern A (across AP vicinity interference) owing to its significance. Remember these patterns occur when APs in an 802.11 infrastructure encounters abrupt endings in succession during short interval of time (i.e. 60 sec). Thus if not controlled, large part of an 802.11 network can suffer from this cascading effect, causing connectivity problems to 802.11 users. Similar to what is reported in the previous work [84], where an entire 802.11 network collapsed due to heavy control, management, and data packet processing required by the arrival of high number of user to APs, each time APs started accepting connections after recovering from a halt condition.

6.6 Online Detection of Anomaly-Related Patterns

6.6.1 Methodology

In section 6.3 we used an offline detection approach with relational databases and standard query language (SQL) in depicting the anomaly-related patterns of 802.11 AP usage investigated in this chapter. In this section, we propose to use the Esper engine [189] for complex event processing and analysis to implement online detection of these patterns. The Esper engine works like a relational database turned upside-down, i.e. instead of storing the data and running queries, the Esper engine allows applications to store queries and run the data through the queries. Response is quick when conditions that match queries occur [189]. Event streams (infinite set of events which are further correlated) considered over a time window period can be highly meaningful and reacting to them quickly is critical for effective management decisions [190].

We take advantage of the continuous execution model provided by the Esper engine to implement online event driven detection and characterization of abrupt endings and their respective anomaly-related patterns. The architecture of our online detection tool is shown in figure 27. We use timestamp and AP IP address associated to each user session ending (i.e. RADIUS log event STOP) to form an input data stream of session endings. From this input event stream, we automatically detect abrupt ending events using the definition provided in chapter 4 (i.e. timestamp where three or more session endings are observed in the same AP) and consequently generate another data stream of abrupt ending events. We use relationship between these two event data streams observed in specified window periods to characterize abrupt endings into appropriate pattern of AP-related anomaly. We assess the impact of using different time window settings (sliding window) for each pattern against detection capability.

Our overall goal is to implement a system that can automatically detect all AP-related anomalous patterns investigated in this chapter as quick as possible, and to promptly raise alarms. Alarms may increase situational awareness to network administrators who can check 802.11 network and act when needed. In this manner, 802.11 administrators may be relieved from the burden of continuous monitoring 802.11 large-scale infrastructures.

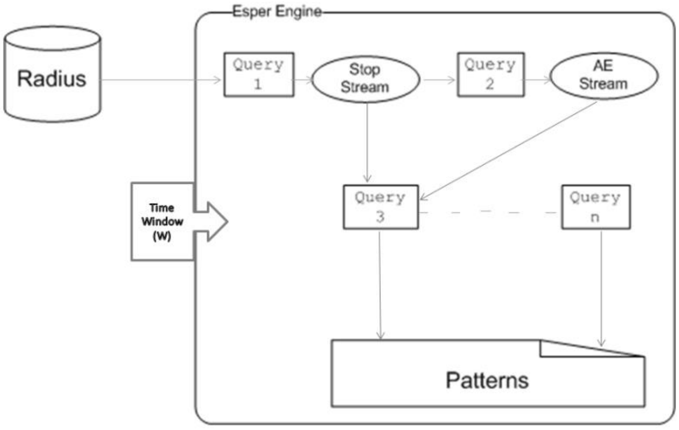


Figure 27. Online system's architecture for anomaly-related patterns detection

Anomaly-Related Pattern Definition: The threshold settings we use here for some patterns detection are slightly different from those used in section 6.3. This is because with our online implementation we aim to reduce detection time yet achieving almost comparable results to the offline detection, which is our baseline for comparison. We employ simple rules for online detection of each anomaly-related pattern on the same data set used in section 6.3. For example: 1) Pattern A: interference across AP vicinity is detected when abrupt ending events occur in a 60 seconds time window between several APs. This threshold is the same for offline detection, because by reducing it there is the danger of splitting one vicinity segment into many segments, on the other hand, further reduction may possibly fade out this pattern completely; 2) Pattern B: AP Persistent interference is detected when four or more abrupt ending events occur in one AP within a 12 hours' time window (our offline detection use 24 hours detection period); 3) Pattern C: AP overload is detected when continued sessions are observed within 1 minute window after abrupt ending event (offline detection use 15 minutes); 4) Pattern D: AP halt/crash is detected when no user session is observed within 5 minutes window after abrupt ending event (offline detection uses 30 minutes); 5) Pattern E: AP Interference is detected if an abrupt ending event is not associated to any of the above mentioned patterns.

We did not investigate user-related patterns F and G, because event streams based on RADIUS log event STOP alone are not enough for characterizing these patterns. These patterns require user information (e.g. User ID) beside each log event STOP, and incorporating it (i.e. user ID event stream) may increase complexity of the online implementation, we leave out this for future work.

6.6.2 Interference Across AP Vicinity Patterns

We test different thresholds of time intervals using sliding window approach while counting the number of abrupt ending events belonging to this particular pattern at various window sizes. Figure 28 shows that as the time window increases the total number of AP vicinity segments and total number of interfering APs increases, with insignificant change noted after 60 seconds; similar to the insights of figure 24. We therefore keep 60 seconds window size as an optimal detection threshold for this pattern, in the same way as in the offline detection of section 6.3.

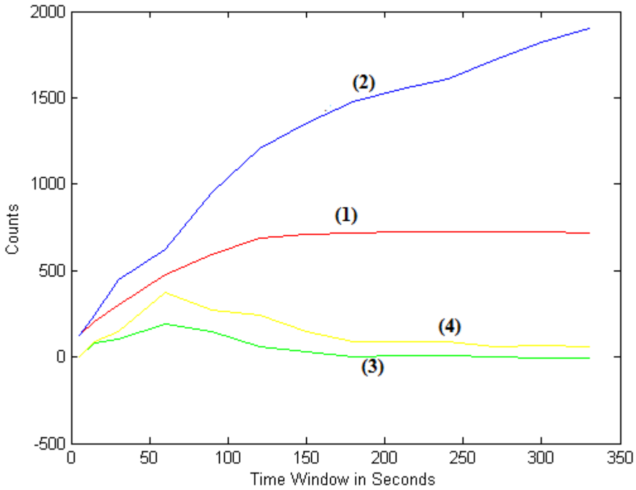
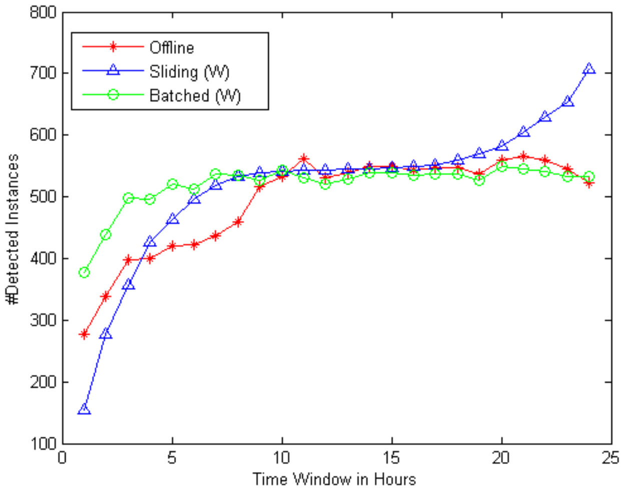


Figure 28. Different time windows tested vs. counts: (1) total number of vicinity segments detected, (2) total number of interfering APs within vicinity segments, (3) difference in number of segments from previous time window, and (4) difference in number of interfering APs between consecutive segments

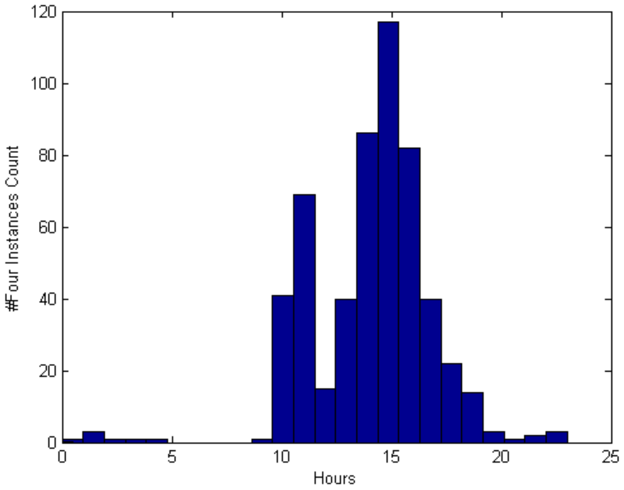
We performed sensitivity analysis on the results obtained from online detection of these patterns (figure 28) against our results of offline detection (figure 24). Sensitivity analysis in this case has maximum 100%, since similar results were obtained. This means our online detector has ability to produce 100% matching results to offline detection using the same 60 seconds threshold. We did not attempt to reduce this threshold further, because further reduction is likely to split abrupt endings that belong to the same vicinity segment into various distinct vicinity segments; while further reduction is likely to destroy vicinity segments completely, as the majority of abrupt endings would appear individually without forming any pattern.

6.6.3 AP Persistent Intereference Patterns

We count the number of abrupt ending events per AP greater or equal to four (similar to offline detection in section 6.3.2) exploiting both sliding and batched windowing approaches at distinct time window intervals. Sliding window's approach means each latest abrupt ending event is checked against the previous occurrence of the abrupt ending event for the same AP within the duration of the chosen window size e.g. in our case 12 hours. While batched window uses fixed intervals of window size e.g. 12 hours, meaning counts of abrupt ending events for all APs is executed at the end of each window size. Batched window is very similar in concept to our offline detection. Figure 29 depicts detection results for different time windows tested using both approaches.



(a) Time windows tested for online detection of AP persistence interference



(b) Hour of the day where 4 abrupt ending events per AP occurred

Figure 29. Results for online detection of AP persistence interference patterns

Figure 29(a) shows the number of instances of AP persistence interferences detected increase with the increase in window size and the detected instances becomes insignificant at around 10 hours until 18 hours. This indicates the chosen 12 hours can be suitable window size for detecting AP persistent interference using both approaches sliding or batched windows. However, batched window consumes more computational and memory resources than sliding window. In batched window events are collected in batches and processed in batches [191], while in sliding window only recent elements of a stream are more important. This fact justifies the use of sliding window over batched window for online detection of these patterns. Figure 29(b) depicts the hour of the day where APs normally reaches four abrupt ending event counts. Remember, occurrences of four abrupt endings to an AP in a given window period is what defines AP persistent of interference patterns (section 6.2). In addition, figure 29(b) shows that the number of 4 counts for AP abrupt ending events follows hotspot usage pattern behavior, with high counts noted between 10:00 and 20:00 hours. These insights may help 802.11 administrators to determine suitable window sizes for both sliding window detection and batched window detection (if so desired), considering usage patterns of an 802.11 network.

We performed sensitivity analysis on the results obtained from our online detector (figure 29(a)) to the results of offline detection (table 14). At 50% reduction in detecting time (i.e. from 24 hours to 12 hours), sensitivity analysis yielded 0.79 true positive rate and 0.11 false positive rate. These means employing 12 hours detection time, we stand a probability of detecting 0.21 wrong (i.e. 21% of wrongly detecting AP persistence interference patterns), and a probability of 0.89 being correct (i.e. 89% of correctly detecting AP persistence interference patterns).

6.6.4 AP Overload Patterns

We check the presence of continued sessions using sliding window approach in a time interval between 1 to 15 minutes after abrupt ending event is detected. The number of AP overload patterns detected increases with the increase in time window size and reaches maximum at 15 minutes. Since user connections are refreshed every 15 minutes, this threshold could work better for offline detection. Offline detection relies much on the log event ALIVE for detection of these patterns. But using online technique we aim to reduce detection time while attaining comparable results to offline. Henceforth, we chose the minimum 1 minute as adequate time window (threshold) for online detection of these patterns. A STOP log event here is monitored in real time within the duration of one minute after AP abrupt ending event; this was not the choice for offline detection which searched for events (ALIVE and STOP) in the interval of 5, 10, and 15 minutes after AP abrupt ending. This also means instead of waiting 15 minute before AP to be declared overloaded, here we only need to wait one minute. Figure 30 shows the number of overload instances detected at each time window’s tested: note close results at 5, 10, and 15 minutes time windows between online and offline detections.

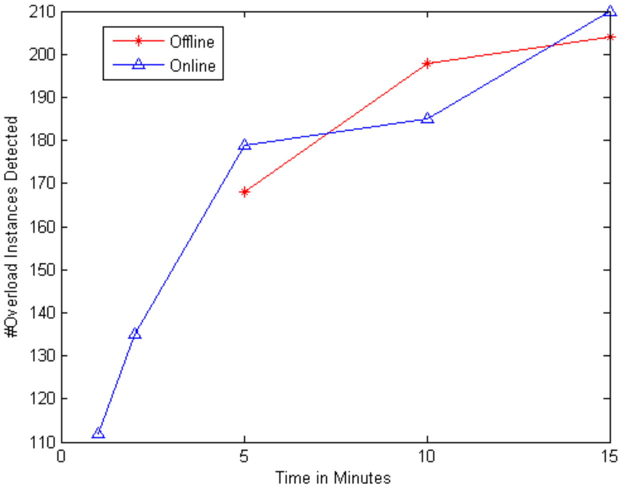


Figure 30. Time windows tested for online detection of AP overload patterns vs. counts

To get an idea of how good our online detection is, we performed sensitivity analysis on the results obtained from using online detector (figure 30) to the results of offline detection (table 15). At 93% reduction in detecting time (i.e. from 15 minutes to 1 minute), sensitivity analysis resulted into 0.87 true positive rate and 0.09 false positive rate. These results means by using 1 minute detection time rather 15 minutes, we can have a probability of detecting 0.13 wrong (i.e. 13% of wrongly detecting AP overload patterns), and a probability of 0.91 being correct (i.e. 91% of correctly detecting AP overload patterns).

6.6.5 AP Halt/Crash Patterns

Similar to section 6.3.4, here we check for APs that do not record any event after abrupt ending using sliding window in the time interval between 5 and 30 minutes. We observe the number of AP halt/crash instances remains the same between 1 and 5 minutes (figure 31). When we increase the window sizes the number of crash instances detected goes to a peak and drops sharply after 15 minutes. This insight indicates that within 5 minutes after abrupt ending event an AP can be pronounced as crashed rather than waiting for the entire 15 minutes or 30 minutes, as in the offline case. However, amid such short detection time slight variations in results is obvious (see figure 31), because offline detection ensures completely no activity is seen to an AP for the entire 30 minutes after abrupt ending, while online detection leaves the room for events (START, ALIVE, and STOP) beyond 5 minutes.

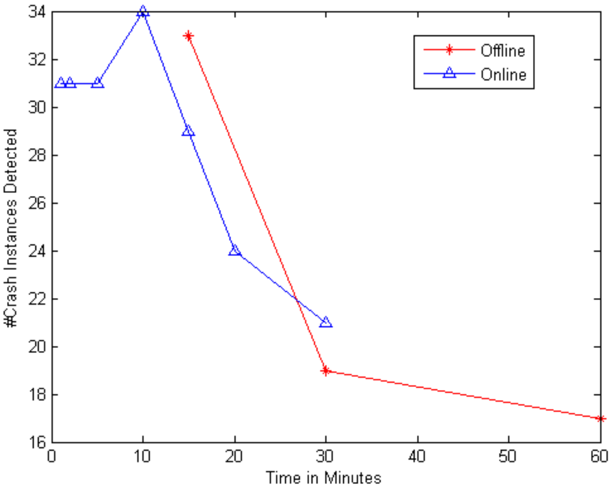


Figure 31. Time windows tested for online detecting of AP halt/crash patterns vs. counts

We also performed sensitivity analysis on the results obtained from our online detector (figure 31) in comparison to the results of offline detection (table 16). At 83% reduction in detecting time (i.e. from 30 minutes to 5 minutes), sensitivity analysis yielded 0.89 true positive rate and 0.09 false positive rate. These means by using 5 minute rather than 30 minutes detection time there is a probability of 0.11 detecting wrong (i.e. 11% of wrongly detecting AP halt/crash patterns), and a probability of 0.91 detecting correct (i.e. 91% of correctly detecting AP halt/crash patterns). These results reveal a very significant gain of using our online detector for the detection of AP halt/crash patterns as compared to offline approach.

6.6.6 AP Interferences Patterns

We follow similar rule as in section 6.3.5 for online detection of these patterns, i.e. a default case of any AP abrupt ending occurrences is interference. We obtain these patterns after removing first patterns related to AP vicinity interference, AP overload, AP halt/crash, and AP persistence interference. With our online implantation, in the beginning any AP abrupt ending occurrence that is detected is first assigned to AP interference patterns (pattern E). If other abrupt endings happened to other APs during 60 seconds, it is then changed and assigned to vicinity pattern (pattern A). If not it is checked for continuing session(s), if found then it is assigned to AP overload pattern (pattern C). If not it is checked for absence of continuing sessions within 5 minutes time interval, if none found then it is assigned to AP halt/crash patterns (pattern D). If all these test fails, it is then remain with its initial assignment which is interference pattern (pattern E). This will only change to AP persistent interference (pattern B) once three or more abrupt ending occurrences are observed for the same AP within 12 hours window period. Otherwise it continues to be pattern E. In this manner, we were able to distinguish automatically between abrupt endings of other patterns from those of AP interference patterns.

Figure 32 presents detection results for different sliding window sizes tested. The insignificant increase in the detected instances of AP interference patterns between 10 and 24 hours matches the insights of figure 29(b). This means around these hours most of the initially assigned AP interference instances (pattern E) have already been reached 4 counts, and thus changed to pattern B (AP persistent interference).

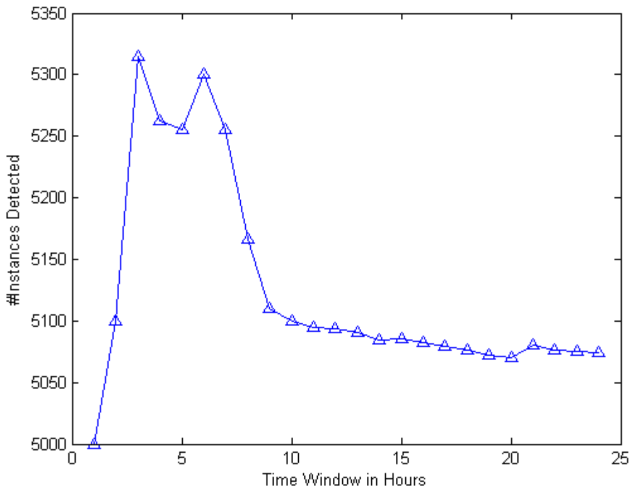


Figure 32. Time windows tested for online detection of AP interference patterns vs. counts

6.6.7 Summary

In this section, we evaluated our proposed online anomaly detector in comparison to the offline algorithm of section 6.2. We presented experimental results for AP-related anomalous patterns evaluated under different thresholds of time windows. Results indicate our online implementation in overall produced matching results to the offline algorithm, specifically with reduced detection time and high rate of true positive and low rate of false positive. For example: 1) for pattern A we obtained similar results, thus sensitivity analysis had maximum 100%; 2) for pattern B we reduced detection time by 50% while achieving about 0.79 true positive rate and 0.11 false positive rate; and 3) for pattern C and D we reduced detection time to over 83% with true positive rate of about 0.87 and false positive of 0.09. The high rates of correctly detected patterns and the significant reduction in detection time, makes our online implementation to be a viable practical and workable approach to detect abrupt ending of connections and their respective anomaly-related patterns in these large-scale 802.11 deployments.

6.7 Conclusions

Anomalies in 802.11 networks can happen individually to APs or collectively following a pattern. In this case, an anomaly detection method needs to test each individual anomaly and search for other related anomalies in the same AP or different APs, in order to establish a pattern. In this chapter, we proposed a method and an algorithm for automatic detection and characterization of anomaly-related patterns of AP usage resulting from the occurrence of AP abrupt ending. We detected the following AP-related patterns: interference across AP vicinity, AP persistent interference, AP halt/crash, AP overload, and AP interference. We also detected user-related patterns such as user authentication failure and user intermittent connectivity to APs.

In addition, we confirmed the existence of significant statistical relationship between anomaly-related patterns occurrence and aggregate 802.11 network usage, in terms of total number of sessions. We demonstrated that this relationship holds for all anomaly-related patterns investigated, except for AP halt/crash and user authentication failure patterns. We showed also occurrences of these anomaly-related patterns can be modeled using simple linear regression models as well as simple exponential distribution models.

We crosschecked the existence of abrupt endings and their respective anomaly-related patterns on different 802.11 hotspots. Our experimental results indicated abrupt endings and their respective anomaly-related patterns are not specific to our environment and also happened in campuses other than ours. This means that our observations, methods, algorithm, and models can be applied across different 802.11 networks. We further presented evaluation of our online implementation in comparison to the offline algorithm. Sensitivity analysis indicated our online detection tool have detected correctly almost all AP-related anomalous patterns investigated in this chapter, with significant reduction in detection time.

Chapter 7

Conclusions and Future Work

The popularity of 802.11 networks and the prolific spread of 802.11-enabled devices have made usage modeling and anomaly detection important aspects for network administrators in maintaining functioning of these networks and in their quest to resolve 802.11 users' problems. While small 802.11 deployments can benefit from traditional management techniques (e.g. protocol analyzer, SNMP polling, human knowledge), large-scale 802.11 deployments require additional management techniques namely, models and algorithm learned from the collected underlying 802.11 network usage data. With available 802.11 usage data set that spans seconds, minutes, hours, days, weeks, months and years plus proper methods, models and algorithms, network administrators can efficiently detect anomalies and plan for usage and capacity of the large-scale 802.11 deployments. Due to lack of effective diagnostic tools and models, in this thesis we contributed with a framework for usage modeling and anomaly detection in the large scale 802.11 network, based on the collected 802.11 usage data.

Contrary to most previous works in modeling wireless sites, we proposed and evaluated different probabilistic models for characterizing access point (AP) usage, including time-independent and time-dependent models that consider week structure usage of AP. We focused on daily counts of keep-alive events that mobile devices generate every 15 minutes while they are connected to the wireless network, rather than user mobility, registration, dwelling and encounter patterns or throughput at AP. We trained and evaluated our models based on data collected at Porto hotspot of Eduroam, the European academic wireless network, and we provided standard cross-validation comparison of models using both the log-likelihood and AIC values on the specific training and test data sets.

In addition, we proposed simple techniques for automatic detection and characterization of patterns of anomaly in the usage of APs. Unlike most proposed approaches in the literature which require instrumenting devices (clients and APs), or introduction of third party devices such hardware sensors, sniffers, and controllers for anomaly detection, in our framework we rely on readily available data and focused on our proposed AP usage pattern "abrupt ending" that happens when APs drop all or significant part of their user connections in a one second window. We proposed simple methodologies, an algorithm, and models for automatic detection and characterization of different anomaly-related patterns associated to AP abrupt ending. Patterns such AP halt/crash, AP overload, AP interference, interference across AP vicinity, AP persistent interference and, user authentication failure and user intermittent connectivity to APs were appropriately detected, characterized, and modeled. We confirmed the existence of significant statistical relationship between abrupt ending occurrences and aggregate

802.11 network usage, in terms of total number of sessions. We demonstrated that this relationship is also true for all anomaly-related patterns investigated except for AP halt/crash and user authentication failure patterns, and further can be modeled using simple linear regression models as well as simple exponential distribution models. We also presented an online implementation of the detection and characterization of abrupt ending and their associated anomaly-related patterns using Esper engine for complex event processing. We used sensitivity analysis to evaluate the performance of our online implementation in comparison to our offline detection methods. The significant reduction in detection time and the high rates of correctly detected patterns indicated that our online implementation can be a viable practical approach to automatically detect AP abrupt ending of connections and their respective anomaly-related patterns in these large-scale 802.11 deployments.

Overall, our study in this thesis presented more insights into the problem of AP usage modeling and anomaly detection. We conclude that significant improvements in AP usage modeling capability can be observed by considering 1) simple time dependency, 2) week-days/week-ends usage structure, and 3) individual day's usage; whereas extending the complexity of time dependency ordering through more previous AP usage samples did not show significant improvements, particularly for daily event count models. Moreover, our results indicated abrupt ending of AP connections and their respective anomaly-related patterns are not environment specific i.e. phenomena beyond our campus, and are in fact the consequences of 1) interference across the 802.11 infrastructure, 2) 802.11 network usage and usage pattern behavior of the AP, also 3) misconfiguration and bugs on the part of the 802.11 AP. Nevertheless, we found no evidence implicating users and specifics of their devices with AP abrupt ending of connections and their resulting patterns. Abrupt endings and their respective anomaly-related patterns appear to be general phenomena in these 802.11 deployments. This means our observations, algorithm, methods, and models can be adopted broadly across different 802.11 networks, with slight change in parameters depending on the underlying 802.11 environment.

We believe insights presented in this thesis reflect accurate assessment of 802.11 AP usage characteristics and can potentially lead to development of better management techniques suitable for the large-scale 802.11 networks. In future work, we plan to enhance our AP usage models to couple with the models for number of users and traffic at APs, also to consider commonalities between APs and detailed distributions of AP usage within a day. We plan to derive network traffic patterns from our AP usage models to use, for example, in network simulations of different applications that may possibly run on 802.11 hotspot infrastructures such as the Eduroam campuses. We also intend to evaluate our detection algorithm looking into other aspects of 802.11 AP usage such as traffic. We plan to further evaluate our online detection tool using live data from our 802.11 networks and to subsequently provide instant feedback to 802.11 network administrators.

References

[1] Gast, M. (2005). 802.11 wireless networks: the definitive guide. O'Reilly Media, Inc.

[2] Stuber, G. L. (1997). Principles of Mobile Communications. Kluwer, Academic Press.

[3] Bianchi, G., Fratta, L., & Oliveri, M. (1996). Performance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LANs. *In Personal, Indoor and Mobile Radio Communications, PIMRC'96., Seventh IEEE International Symposium on (Vol. 2, pp. 392-396)*. IEEE.

[4] Gummalla, A. C. V., & Limb, J. O. (2000). Wireless medium access control protocols. *Communications Surveys & Tutorials*, IEEE, 3(2), 2-15.

[5] Xu, S., & Saadawi, T. (2002). Revealing the problems with 802.11 medium access control protocol in multi-hop wireless ad hoc networks. *Computer Networks*, 38(4), 531-548.

[6] Bianchi, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function. *Selected Areas in Communications, IEEE Journal on*, 18(3), 535-547.

[7] Bing, B. (1999). Measured performance of the IEEE 802.11 wireless LAN. *In Local Computer Networks, LCN'99. Conference on (pp. 34-42).* IEEE.

[8] T. D., Niemi, M., & Saarinen, J. (2002). Trends in personal wireless data communications. *Computer Communications*, 25(1), 84-99.

[9] Stallings, W. (2004). IEEE 8O2. 11: wireless LANs from a to n. IT professional, 6(5), 32-37.

[10] Nguyen, D. N., & Krunz, M. (2013). Cooperative MIMO in wireless networks: recent developments and challenges. *Network, IEEE,* 27(4).

[11] Pathak, P. H., & Dutta, R. (2011). A survey of network design problems and joint design approaches in wireless mesh networks. *Communications Surveys & Tutorials, IEEE,* 13(3), 396-428.

[12] Patras, P., Qi, H., & Malone, D. (2014). Mitigating collisions through power-hopping to improve 802.11 performance. *Pervasive and Mobile Computing*, 11, 41-55.

[13] Xu, F., Zhu, X., Tan, C., Li, Q., Yan, G., & Wu, J. (2013). Smartassoc: Decentralized access point selection algorithm to improve throughput. *Parallel and Distributed Systems, IEEE Transactions on*, 24 (12), 2482 – 2491.

[14] Dargie, W., & Schill, A. (2011). Stability and performance analysis of randomly deployed wireless networks. *Journal of Computer and System Sciences*, 77(5), 852-860.

[15] Leinwand, A., & Conroy, K. F. (1996). Network management: a practical perspective. Unix and Open Systems Series. Reading, MA: Addison-Wesley, 2nd ed.

[16] Klerer, S. M. (1988). The OSI management architecture: an overview. *Network, IEEE,* 2(2), 20-29.

[17] Simoneau, P. (1999). SNMP network management. McGraw-Hill, Inc..

[18] Lee, J. S., & Hsu, P. L. (2004). Design and implementation of the SNMP agents for remote monitoring and control via UML and Petri nets. *Control Systems Technology, IEEE Transactions on*, 12(2), 293-302.

[19] Caruso, R. E. (1990). Network management: a tutorial overview. *Communications Magazine, IEEE,* 28(3), 20-25.

[20] Hassan, R., Razali, R., Mohseni, S., Mohamad, O., & Ismail, Z. (2009). Architecture of Network Management Tools for Heterogeneous System. *International Journal of Computer Science and Information Security, (IJCSIS),* Vol. 6, No. 3.

[21] Freeman, B. D. (2010). Network Configuration Management. *In Guide to Reliable Internet Services and Applications* (pp. 255-275). Springer London.

[22] Goyal, P., Mikkilineni, R., & Ganti, M. (2009). FCAPS in the business services fabric model. In Enabling Technologies: *Infrastructures for Collaborative Enterprises, WETICE'09. 18th IEEE International Workshops on (pp. 45-51).* IEEE.

[23] Lu, X., Zhou, W., & Song, J. (2010). Key issues of future network management. In *Computer Application and System Modeling (ICCASM), International Conference on (Vol. 11, pp. V11-649)*. IEEE.

[24] Said, S. B. H., Guillouard, K., & Bonnin, J. M. (2013). A Comparative Study on Security Implementation in EPS/LTE and WLAN/802.11. In *Wireless Networks and Security* (pp. 457-489). Springer Berlin Heidelberg.

[25] Pras, A., Drevers, T., van de Meent, R., & Quartel, D. (2004). Comparing the performance of SNMP and web services-based management. *Network and Service Management, IEEE Transactions on*, 1(2), 72-82.

[26] Pan, H., Li, T., & Shi, Y. (2014). Computer Network Monitoring System Based on Information Processing Technology. In *Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 1* (pp. 721-728). Springer Berlin Heidelberg.

[27] Surputheen, M. M., Ravi, G., & Srinivasan, R. (2012). SNMP Based Network Optimization Technique Using Genetic Algorithms. *International Journal of Computer Science Issues (IJCSI),* 9(2).

[28] Pavlou, G., Flegkas, P., Gouveris, S., & Liotta, A. (2004). On management technologies and the potential of web services. *Communications Magazine, IEEE,* 42(7), 58-66.

[29] Shang-Fu, G., & Xiao-Li, Y. (2012). Study and Design of Integrated Transmission Network Management System Based on CORBA and Web. In *Industrial Control and Electronics Engineering (ICICEE), 2012 International Conference on (pp. 600-603).* IEEE.

[30] Gupta, A. (2006). Network management: Current trends and future perspectives. *Journal of Network and Systems Management*, 14(4), 483-491.

[31] Satoh, I. (2002). A framework for building reusable mobile agents for network management. In *Network Operations and Management Symposium, 2002. NOMS 2002. 2002 IEEE/IFIP (pp. 51-64).* IEEE.

[32] Yang, S. Y., & Chang, Y. Y. (2011). An active and intelligent network management system with ontology-based and multi-agent techniques. *Expert Systems with Applications,* 38(8), 10320-10342.

[33] Verma, D. C. (2002). Simplifying network administration using policy-based management. *Network, IEEE,* 16(2), 20-26.

[34] Samaan, N., & Karmouch, A. (2009). Towards autonomic network management: an analysis of current and future research directions. *Communications Surveys & Tutorials, IEEE,* 11(3), 22-36.

[35] RFC 2865 (RADIUS) protocol, http://tools.ietf.org/html/rfc2865 last accessed November 2014.

[36] RFC 2866 RADIUS acccounting, http://tools.ietf.org/html/rfc2866 last accessed November 2014.

[37] CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. http://crawdad.cs.dartmouth.edu/index.php last accessed November 2014.

[38] MobiLib: Community-wide Library of Mobility and Wireless Networks Measurements. http://nile.usc.edu/MobiLib last accessed November 2014.

[39] Allahdadi, A., Morla, R., Aguiar, A., & Cardoso, J. S. (2013). Predicting short 802.11 sessions from RADIUS usage data. In *Local Computer Networks Workshops (LCN Workshops), IEEE 38th Conference on (pp. 1-8).* IEEE.

[40] Baras, K., & Moreira, A. (2010). Anomaly detection in university campus WiFi zones. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 8th IEEE International Conference on (pp. 202-207).* IEEE.

[41] Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26, 301-313.

[42] H Zang, H., & Bolot, J. C. (2007). Mining Call Data to Increase the Robustness of Cellular Networks to Signaling DoS Attacks. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking (pp. 123-134)*. ACM.

[43] Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *Communications Magazine, IEEE,* 48(9), 140-150.

[44] Zhu, Y., Liu, X., & Wang, Y. (2013). Pervasive Urban Sensing with Large-Scale Mobile Probe Vehicles, *International Journal of Distributed Sensor Networks* 762503,

[45] González, M. C., Hidalgo, C. A., & Barabási, A. L. (2009). Understanding individual human mobility patterns. *Nature,* 458(7235), 238-238.

[46] Brockmann, D., & Theis, F. (2008). Money circulation, trackable items, and the emergence of universal human mobility patterns. *Pervasive Computing, IEEE,* 7(4), 28-35.

[47] Massa, D., & Morla, R. (2013). Modeling 802.11 AP usage through daily keep-alive event counts. *Wireless networks,* 19(5), 1005-1022.

[48] Massa, D., & Morla, R. (2010). Modeling 802.11 AP Usage through Daily Keep-Alive Event Counts. In *16th International Conference on Network-Based Information Systems (pp. 195-200).* IEEE.

[49] Massa, D., & Morla, R. (2013). Abrupt ending of 802.11 ap connections. In *Computers and Communications (ISCC), IEEE Symposium on* (pp. 000348-000353). IEEE.

[50] Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender Systems Handbook* (pp. 39-71).

[51] Tang, D., & Baker, M. (2000). Analysis of a local-area wireless network. In *Proceedings of the 6th annual international conference on Mobile computing and networking* (pp. 1-10). ACM.

[52] Tang, D., & Baker, M. (2002). Analysis of a metropolitan-area wireless network. *Wireless Networks, 8*(2-3), 107-120.

[53] Balachandran, A., Voelker, G. M., Bahl, P., & Rangan, P. V. (2002). Characterizing user behavior and network performance in a public wireless LAN. In *ACM SIGMETRICS Performance Evaluation Review* (Vol. 30, No. 1, pp. 195-205). ACM.

[54] Kotz, D., & Essien, K. (2002). Analysis of a Campus-wide Wireless Network. In *In Proceedings of ACM Mobicom*. ACM.

[55] Henderson, T., Kotz, D., & Abyzov, I. (2004). The changing usage of a mature campus-wide wireless network. In *Proceedings of the 10th annual international conference on Mobile computing and networking* (pp. 187-201). ACM.

[56] Balazinska, M., & Castro, P. (2003). Characterizing mobility and network usage in a corporate wireless local-area network. In *Proceedings of the 1st international conference on Mobile systems, applications and services* (pp. 303-316). ACM.

[57] Kim, M., & Kotz, D. (2007). Periodic properties of user mobility and access-point popularity. *Personal and Ubiquitous Computing*, *11*(6), 465-479.

[58] Kim, M., & Kotz, D. (2005). Modeling users' mobility among WiFi access points. In *Papers presented at the 2005 workshop on Wireless traffic measurements and modeling* (pp. 19-24). USENIX Association.

[59] Boc, M., Fladenmuller, A., & De Amorim, M. D. (2007). Towards self-characterization of user mobility patterns. In *Mobile and Wireless Communications Summit, 2007. 16th IST* (pp. 1-5). IEEE.

[60] Kim, M., Kotz, D., & Kim, S. (2006). Extracting a Mobility Model from Real User Traces. In *INFOCOM* (Vol. 6, pp. 1-13). IEEE.

[61] McNett, M., & Voelker, G. M. (2005). Access and mobility of wireless PDA users. *ACM SIGMOBILE Mobile Computing and Communications Review*, *9*(2), 40-55.

[62] Hsu, W. J., & Helmy, A. (2006). On modeling user associations in wireless LAN traces on university campuses. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 4th International Symposium on* (pp. 1-9). IEEE.

[63] Chen, Y. C., Kurose, J., & Towsley, D. (2012, March). A mixed queueing network model of mobility in a campus wireless network. In *INFOCOM, Proceedings* (pp. 2656-2660). IEEE.

[64] Ghosh, J., Beal, M. J., Ngo, H. Q., & Qiao, C. (2006). On profiling mobility and predicting locations of wireless users. In *Proceedings of the 2nd international workshop on Multi-hop ad hoc networks: from theory to reality* (pp. 55-62). ACM.

[65] Song, L., Kotz, D., Jain, R., & He, X. (2006). Evaluating next-cell predictors with extensive Wi-Fi mobility data. *Mobile Computing, IEEE Transactions on, 5*(12), 1633-1649.

[66] Kim, J., & Helmy, A. (2011). The evolution of wlan user mobility and its effect on prediction. In *Wireless Communications and Mobile Computing Conference (IWCMC), 7th International* (pp. 226-231). IEEE.

[67] Lee, J. K., & Hou, J. C. (2006). Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application. In *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing* (pp. 85-96). ACM.

[68] Chon, Y., Shin, H., Talipov, E., & Cha, H. (2012). Evaluating mobility models for temporal prediction with high-granularity mobility data. In *Pervasive computing and communications (percom), IEEE international conference on* (pp. 206-212). IEEE.

[69] Gao, W., & Cao, G. (2010). Fine-grained mobility characterization: steady and transient state behaviors. In *Proceedings of the eleventh ACM international symposium on Mobile ad hoc networking and computing* (pp. 61-70). ACM.

[70] Abu-Ghazaleh, H., & Alfa, A. S. (2010). Application of mobility prediction in wireless networks using markov renewal theory. *Vehicular Technology, IEEE Transactions on*, *59*(2), 788-802.

[71] Hong, J., & Kim, H. (2011). An empirical framework for user mobility models: Refining and modeling user registration patterns. *Journal of Computer and System Sciences*, 77(5), 869-883.

[72] Hong, J., & Kim, H. (2013). A Dual Mobility Model with User Profiling: Decoupling User Mobile Patterns from Association Patterns. *The Computer Journal*, *56*(6), 771-784.

[73] Zhao, M., & Wang, W. (2009). A unified mobility model for analysis and simulation of mobile wireless networks. *Wireless Networks*, *15*(3), 365-389.

[74] Resta, G., & Santi, P. (2008). WiQoSM: An integrated QoS-aware mobility and user behavior model for wireless data networks. *IEEE Transactions on Networking Mobile Computing*, 7(2), 187–198.

[75] Lee, K., Hong, S., Kim, S. J., Rhee, I., & Chong, S. (2009). Slaw: A new mobility model for human walks. In *INFOCOM*, (pp. 855-863). IEEE.

[76] Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., & Chong, S. (2011). On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19(3), 630-643.

[77] Kosta, S., Mei, A., & Stefa, J. (2010). Small world in motion (SWIM): Modeling communities in ad-hoc mobile networking. In *Sensor Mesh and Ad Hoc Communications and Networks (SECON), 27th Annual IEEE Communications Society Conference on* (pp. 1-9). IEEE.

[78] Meneses, F., & Moreira, A. (2012). Large scale movement analysis from WiFi based location data. In *Indoor Positioning and Indoor Navigation (IPIN), International Conference on*. IEEE.

[79] Balasubramanian, A., Mahajan, R., Venkataramani, A., Levine, B. N., & Zahorjan, J. (2008). Interactive wifi connectivity for moving vehicles. *ACM SIGCOMM Computer Communication Review*, *38*(4), 427-438.

[80] Mahajan, R., Zahorjan, J., & Zill, B. (2007). Understanding WiFi-based connectivity from moving vehicles. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 321-326). ACM.

[81] Hsu, W.-j., & Helmy, A. (2010). “On nodal encounter patterns in wireless LAN traces”, *IEEE Transaction on Mobile Computing*, 9(11), 1563–1577.

[82] Hsu, W. J., Dutta, D., & Helmy, A. (2007). Mining behavioral groups in large wireless LANs. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking* (pp. 338-341). ACM.

[83] Hsu, W. J., Dutta, D., & Helmy, A. (2012). Structural analysis of user association patterns in university campus wireless lans. *Mobile Computing, IEEE Transactions on*, *11*(11), 1734-1748.

[84] Hsu, W. J., Spyropoulos, T., Psounis, K., & Helmy, A. (2007). Modeling time-variant user mobility in wireless mobile networks. In *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE* (pp. 758-766). IEEE.

[85] Hsu, W. J., Spyropoulos, T., Psounis, K., & Helmy, A. (2009). Modeling spatial and temporal dependencies of user mobility in wireless mobile networks. *IEEE/ACM Transactions on Networking (TON)*, 17(5), 1564-1577.

[86] Chen, Y. C., Rosensweig, E., Kurose, J., & Towsley, D. (2010). Group detection in mobility traces. In *Proceedings of the 6th international wireless communications and mobile computing conference* (pp. 875-879). ACM.

[87] Karagiannis, T., Le Boudec, J. Y., & Vojnovic, M. (2010). Power law and exponential decay of intercontact times between mobile devices. *Mobile Computing, IEEE Transactions on*, *9*(10), 1377-1390.

[88] Williams, M. J., Whitaker, R. M., & Allen, S. M. (2012). Measuring individual regularity in human visiting patterns. In *International Confernece on Social Computing (SocialCom)* (pp. 117-122). IEEE.

[89] Papadopouli, M., Shen, H., & Spanakis, M. (2005). Modeling client arrivals at access points in wireless campus-wide networks. In *14th IEEE Workshop on Local and Metropolitan Area Networks* (pp. 18-21). IEEE.

[90] Ghosh, A., Jana, R., Ramaswami, V., Rowland, J., & Shankaranarayanan, N. K. (2011). Modeling and characterization of large-scale Wi-Fi traffic in public hot-spots. In *INFOCOM, Proceedings* (pp. 2921-2929). IEEE.

[91] Jain, R., Lelescu, D., & Balakrishnan, M. (2007). Model T: a model for user registration patterns based on campus WLAN data. *Wireless Networks*, *13*(6), 711-735.

[92] Chinchilla, F., Lindsey, M., & Papadopouli, M. (2004). Analysis of wireless information locality and association patterns in a campus. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies* (Vol. 2, pp. 906-917). IEEE.

[93] Lelescu, D., Kozat, U. C., Jain, R., & Balakrishnan, M. (2006). Model T++: an empirical joint space-time registration model. In *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing* (pp. 61-72). ACM.

[94] Zola, E., & Barcelo-Arroyo, F. (2012). Distribution of the frequency of connections in academic WLAN networks. In *ICNS 2012, The Eighth International Conference on Networking and Services* (pp. 69-74).

[95] Zola, E., & Barcelo-Arroyo, F. (2013). Characterizing User Behavior in a European Academic WiFi Network. *International Journal of Handheld Computing Research (IJHCR)*, *4*(2), 55-68.

[96] Papadopouli, M., Shen, H., & Spanakis, M. (2005). Characterizing the duration and association patterns of wireless access in a campus. In *Wireless Conference 2005-Next Generation Wireless and Mobile Communications and Services (European Wireless), 11th European* (pp. 1-7). VDE.

[97] Mahanti, A., Williamson, C., & Arlitt, M. (2007). Remote analysis of a distributed WLAN using passive wireless-side measurement. *Performance Evaluation*, *64*(9), 909-932.

[98] Phillips, C., & Singh, S. (2008). An empirical activity model for wlan users. In *INFOCOM, 08. The 27th Conference on Computer Communications*. IEEE.

[99] Zola, E., & Barcelo-Arroyo, F. (2009). Impact of mobility models on the cell residence time in WLAN networks. In *Sarnoff Symposium, SARNOFF'09 (pp. 1-5).* IEEE.

[100] Manweiler, J., Santhapuri, N., Choudhury, R. R., & Nelakuditi, S. (2013). Predicting length of stay at WiFi hotspots. In *INFOCOM, 13 Proceedings (pp. 3102-3110).* IEEE.

[101] Papadopouli, M., Shen, H., Raftopoulos, E., Ploumidis, M., & Hernandez-Campus, F. (2005). Short-term traffic forecasting in a campus-wide wireless network. In *Personal, Indoor and Mobile Radio Communications, PIMRC, 05. 16th International Symposium on* (Vol. 3, pp. 1446-1452). IEEE.

[102] Tzagkarakis, G., Papadopouli, M., & Tsakalides, P. (2009). Trend forecasting based on Singular Spectrum Analysis of traffic workload in a large-scale wireless LAN. *Performance Evaluation,* 66(3), 173-190.

[103] Kulkarni, P., Lewis, T., & Fan, Z. (2011). Simple traffic prediction mechanism and its applications in wireless networks. *Wireless Personal Communications*, 59(2), 261-274.

[104] Hernandez-Campos, F., & Papadopouli, M. (2005). A comparative measurement study the workload of wireless access points in campus networks. In *Personal, Indoor and Mobile Radio Communications, PIMRC, 05. 16th International Symposium on (Vol. 3, pp. 1776-1780)*. IEEE.

[105] Hernández-Campos, F., Karaliopoulos, M., Papadopouli, M., & Shen, H. (2006). Spatio-temporal modeling of traffic workload in a campus WLAN. In *Proceedings of the 2nd annual international workshop on Wireless internet* (p. 1). ACM.

[106] Karaliopoulos, M., Papadopouli, M., Raftopoulos, E., & Shen, H. (2007). On scalable measurement-driven modeling of traffic demand in large WLANs. In *Local & Metropolitan Area Networks, LANMAN, 07. 15th Workshop on* (pp. 102-110). IEEE.

[107] Meng, X. G., Wong, S. H., Yuan, Y., & Lu, S. (2004). Characterizing flows in large wireless data networks. In *Proceedings of the 10th annual international conference on Mobile computing and networking* (pp. 174-186). ACM.

[108] Chlebus, E., & Divgi, G. (2007). The Pareto or truncated Pareto distribution? Measurement-based modeling of session traffic for Wi-Fi wireless Internet access. In *Wireless Communications and Networking Conference, WCNC, 07* (pp. 3625-3630). IEEE.

[109] Mulhanga, M. M., Lima, S. R., & Carvalho, P. (2011). Characterising Eduroam WLANs Usage Trends: A Case Study. In *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 19th International Symposium on* (pp. 447-449). IEEE.

[110] Afanasyev, M., Chen, T., Voelker, G. M., & Snoeren, A. C. (2008). Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement* (pp. 85-98). ACM.

[111] Afanasyev, M., Chen, T., Voelker, G. M., & Snoeren, A. C. (2010). Usage patterns in an urban wifi network. *Networking, IEEE/ACM Transactions on*, 18(5), 1359-1372.

[112] Blinn, D. P., Henderson, T., & Kotz, D. (2005). Analysis of a Wi-Fi hotspot network. In *Papers presented at the workshop on Wireless traffic measurements and modeling* (pp. 1-6). USENIX Association.

[113] Divgi, G., & Chlebus, E. (2013). Characterization of user activity and traffic in a commercial nationwide Wi-Fi hotspot network: global and individual metrics. *Wireless networks*, 19(7), 1783-1805.

[114] Ojala, T., Hakanen, T., Makinen, T., & Rivinoja, V. (2005). Usage analysis of a large public wireless LAN. In *Wireless Networks, Communications and Mobile Computing, International Conference on* (Vol. 1, pp. 661-667). IEEE.

[115] Wamser, F., Pries, R., Staehle, D., Heck, K., & Tran-Gia, P. (2011). Traffic characterization of a residential wireless Internet access. *Telecommunication Systems*, *48*(1-2), 5-17.

[116] Balachandran, A., Voelker, G. M., & Bahl, P. (2005). Wireless hotspots: current challenges and future directions. *Mobile Networks and Applications*, *10*(3), 265-274.

[117] Jardosh, A. P., Ramachandran, K. N., Almeroth, K. C., & Belding-Royer, E. M. (2005). Understanding link-layer behavior in highly congested IEEE 802.11 b wireless networks. In *Proceedings of the ACM SIGCOMM workshop on Experimental approaches to wireless network design and analysis* (pp. 11-16). ACM.

[118] Jardosh, A. P., Ramachandran, K. N., Almeroth, K. C., & Belding-Royer, E. M. (2005). Understanding congestion in IEEE 802.11 b wireless networks. In *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement* (pp. 25-25). USENIX Association.

[119] Raghavendra, R., Belding, E. M., Papagiannaki, K., & Almeroth, K. C. (2010). Unwanted link layer traffic in large IEEE 802.11 wireless networks. *IEEE Transactions on Mobile Computing*, 9(9), 1212-1225.

[120] Acharya, P. A. K., Sharma, A., Belding, E. M., Almeroth, K. C., & Papagiannaki, K. (2008). Congestion-aware rate adaptation in wireless networks: A measurement-driven approach. In *Sensor, Mesh and Ad Hoc Communications and Networks, SECON'08. 5th Annual IEEE Communications Society Conference on* (pp. 1-9). IEEE.

[121] Velayos, H., Más, I., & Karlsson, G. (2006). Overload protection for ieee 802.11 cells. In *Quality of Service, IWQoS 06. 14th IEEE International Workshop on* (pp. 149-158). IEEE.

[122] Abusubaih, M., Wiethoelter, S., Gross, J., & Wolisz, A. (2008). A new access point selection policy for multi-rate IEEE 802.11 WLANs. *International Journal of Parallel, Emergent and Distributed Systems*, *23*(4), 291-307.

[123] Jardosh, A. P., Mittal, K., Ramachandran, K. N., Belding, E. M., & Almeroth, K. C. (2006). IQU: practical queue-based user association management for WLANs. In *Proceedings of the 12th annual international conference on Mobile computing and networking* (pp. 158-169). ACM.

[124] Velayos, H., Aleo, V., & Karlsson, G. (2004). Load balancing in overlapping wireless LAN cells. In *International Conference on Communications* (Vol. 7, pp. 3833-3836). IEEE.

[125] Murty, R., Padhye, J., Chandra, R., Wolman, A., & Zill, B. (2008). Designing High Performance Enterprise Wi-Fi Networks. In *NSDI* (Vol. 8, pp. 73-88).

[126] Nicholson, A. J., Chawathe, Y., Chen, M. Y., Noble, B. D., & Wetherall, D. (2006). Improved access point selection. In *Proceedings of the 4th international conference on Mobile systems, applications and services* (pp. 233-245). ACM.

[127] Bejerano, Y., & Han, S. J. (2009). Cell breathing techniques for load balancing in wireless LANs. *Mobile Computing, IEEE Transactions on, 8*(6), 735-749.

[128] Mittal, K., Belding, E. M., & Suri, S. (2008). A game-theoretic analysis of wireless access point selection by mobile users. *Computer Communications*, *31*(10), 2049-2062.

[129] Pan, H. J., & Keshav, S. (2006). Detection and repair of faulty access points. In *Wireless Communications and Networking Conference, WCNC 06.* (Vol. 1, pp. 532-538). IEEE.

[130] de Deus, F. E., Puttini, R. S., Molinaro, L. F., Abdalla, H., Amvame-Nze, G., & Kabara, J. (2006). Fault tolerance in IEEE 802.11 WLANs. In *Telecommunications Symposium, International* (pp. 626-631). IEEE.

[131] Adya, A., Bahl, P., Chandra, R., & Qiu, L. (2004). Architecture and techniques for diagnosing faults in IEEE 802.11 infrastructure networks. In *Proceedings of the 10th annual international conference on Mobile computing and networking* (pp. 30-44). ACM.

[132] Elhadef, M., Boukerche, A., & Elkadiki, H. (2008). A distributed fault identification protocol for wireless and mobile ad hoc networks. *Journal of Parallel and Distributed Computing*, 68(3), 321-335.

[133] Akella, A., Judd, G., Seshan, S., & Steenkiste, P. (2007). Self-management in chaotic wireless deployments. *Wireless Networks*, 13(6), 737-755.

[134] Broustis, I., Papagiannaki, K., Krishnamurthy, S. V., Faloutsos, M., & Mhatre, V. P. (2010). Measurement-driven guidelines for 802.11 WLAN design. *IEEE/ACM Transactions on Networking (TON)*, *18*(3), 722-735.

[135] Moriuchi, A., Murase, T., Oguchi, M., Baid, A., Sagari, S., Seskar, I., & Raychaudhuri, D. (2013). Measurement study of adjacent channel interference in mobile WLANs. In *Communications Workshops (ICC), International Conference on* (pp. 566-570). IEEE.

[136] Gummadi, R., Wetherall, D., Greenstein, B., & Seshan, S. (2007). Understanding and mitigating the impact of RF interference on 802.11 networks. *ACM SIGCOMM Computer Communication Review*, *37*(4), 385-396.

[137] Shrivastava, V., Rayanchu, S. K., Banerjee, S., & Papagiannaki, K. (2011). PIE in the Sky: Online Passive Interference Estimation for Enterprise WLANs. In *NSDI* (Vol. 11, p. 20).

[138] Lakshminarayanan, K., Seshan, S., & Steenkiste, P. (2011). Understanding 802.11 performance in heterogeneous environments. In *Proceedings of the 2nd ACM SIGCOMM workshop on Home networks* (pp. 43-48). ACM.

[139] Cai, K., Blackstock, M., Feeley, M. J., & Krasic, C. (2009). Non-intrusive, dynamic interference detection for 802.11 networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference* (pp. 377-383). ACM.

[140] Yang, J., Draper, S. C., & Nowak, R. (2012). Passive learning of the interference graph of a wireless network. In *Information Theory Proceedings (ISIT), IEEE International Symposium on* (pp. 2735-2740). IEEE.

[141] Chang, H., Misra, V., & Rubenstein, D. (2006). A General Model and Analysis of Physical Layer Capture in 802.11 Networks. In *INFOCOM*. IEEE.

[142] Qiu, L., Zhang, Y., Wang, F., Han, M. K., & Mahajan, R. (2007). A general model of wireless interference. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking* (pp. 171-182). ACM.

[143] Kashyap, A., Paul, U., & Das, S. R. (2010). Deconstructing interference relations in WiFi networks. In *Sensor Mesh and Ad Hoc Communications and Networks (SECON), 7th Annual IEEE Communications Society Conference on* (pp. 1-9). IEEE.

[144] Paul, U., Kashyap, A., Maheshwari, R., & Das, S. R. (2013). Passive measurement of interference in WiFi networks with application in misbehavior detection. *Mobile Computing, IEEE Transactions on*, *12*(3), 434-446.

[145] Qu, G., & Nefcy, M. M. (2010). RAPiD: An indirect rogue access points detection system. In *Performance Computing and Communications Conference (IPCCC), IEEE 29th International* (pp. 9-16). IEEE.

[146] Kao, K. F., Liao, I., & Li, Y. C. (2009). Detecting rogue access points using client-side bottleneck bandwidth analysis. *Computers and Security*, *28*(3), 144-152.

[147] Nikbakhsh, S., Manaf, A. B. A., Zamani, M., & Janbeglou, M. (2012). A novel approach for rogue access point detection on the client-side. In *Advanced Information Networking and Applications Workshops (WAINA), 26th International Conference on* (pp. 684-687). IEEE.

[148] Ma, L., Teymorian, A. Y., & Cheng, X. (2008). A hybrid rogue access point protection framework for commodity Wi-Fi networks. In *INFOCOM 2008. The 27th Conference on Computer Communications*. IEEE.

[149] Watkins, L., Beyah, R., & Corbett, C. (2007). A passive approach to rogue access point detection. In *Global Telecommunications Conference, GLOBECOM'07.* (pp. 355-360). IEEE.

[150] Venkataraman, A., & Beyah, R. (2009). Rogue access point detection using innate characteristics of the 802.11 mac. In *Security and Privacy in Communication Networks* (pp. 394-416). Springer Berlin Heidelberg.

[151] Han, H., Sheng, B., Tan, C. C., Li, Q., & Lu, S. (2011). A timing-based scheme for rogue AP detection. *Parallel and Distributed Systems, IEEE Transactions on*, 22(11), 1912-1925.

[152] Han, H., Sheng, B., Tan, C. C., Li, Q., & Lu, S. (2009). A measurement based rogue ap detection scheme. In *INFOCOM 09, IEEE* (pp. 1593-1601). IEEE.

[153] Shivaraj, G., Song, M., & Shetty, S. (2010). Using Hidden Markov Model to detect rogue access points. *Security and Communication Networks*, 3(5), 394-407

[154] Rodrig, M., Reis, C., Mahajan, R., Wetherall, D., & Zahorjan, J. (2005). Measurement-based characterization of 802.11 in a hotspot setting. In *Proceedings of the ACM SIGCOMM workshop on Experimental approaches to wireless network design and analysis* (pp. 5-10). ACM.

[155] Raghavendra, R., Belding, E. M., Papagiannaki, K., & Almeroth, K. C. (2007). Understanding handoffs in large ieee 802.11 wireless networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 333-338). ACM.

[156] Cheng, Y. C., Afanasyev, M., Verkaik, P., Benkö, P., Chiang, J., Snoeren, A. C., & Voelker, G. M. (2007). *Automating cross-layer diagnosis of enterprise wireless networks* (Vol. 37, No. 4, pp. 25-36). ACM.

[157] Ergin, M. A., Ramachandran, K., & Gruteser, M. (2008). An experimental study of inter-cell interference effects on system performance in unplanned wireless LAN deployments. *Computer Networks*, 52(14), 2728-2744

[158] Pelletta, E., & Velayos, H. (2005). Performance measurements of the saturation throughput in IEEE 802.11 access points. In *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, WIOPT 2005. Third International Symposium on* (pp. 129-138). IEEE.

[159] Sarkar, N. I., & Lo, E. (2011). Performance studies of 802.11 g for various AP configuration and placement. In *Computers & Informatics (ISCI), IEEE Symposium on* (pp. 29-34). IEEE.

[160] Choi, J., & Shin, K. G. (2012). QoS provisioning for large-scale multi-ap WLANs. *Ad Hoc Networks*, *10*(2), 174-185.

[161] Soroush, H., Gilbert, P., Banerjee, N., Levine, B. N., Corner, M., & Cox, L. (2011). Concurrent Wi-Fi for mobile users: analysis and measurements. In *Proceedings of the Seventh Conference on emerging Networking Experiments and Technologies* (p.4). ACM.

[162] Wang, S., Guo, W., & O'Farrell, T. (2012). Low energy indoor network: deployment optimisation. *EURASIP Journal on Wireless Communications and Networking*, *2012*(1), 1-15.

[163] Judd, G., & Steenkiste, P. (2007). Understanding link-level 802.11 behavior: Replacing convention with measurement. In *Proceedings of the 3rd international conference on Wireless internet* (p. 19). ICST.

[164] Mare, S., Kotz, D., & Kumar, A. (2010). Experimental validation of analytical performance models for IEEE 802.11 networks. In *Communication Systems and Networks (COMSNETS), 2010 Second International Conference on* (pp. 1-8). IEEE.

[165] Patras, P., Qi, H., & Malone, D. (2012). Exploiting the capture effect to improve WLAN throughput. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), IEEE International Symposium on a* (pp. 1-9). IEEE.

[166] Hung, F. Y., & Marsic, I. (2010). Performance analysis of the IEEE 802.11 DCF in the presence of the hidden stations. *Computer Networks,* 54(15), 2674-2687.

[167] Sheth, A., Doerr, C., Grunwald, D., Han, R., & Sicker, D. (2006). MOJO: A distributed physical layer anomaly detection system for 802.11 WLANs. In *Proceedings of the 4th international conference on Mobile systems, applications and services* (pp. 191-204). ACM.

[168] Cheng, Y. C., Bellardo, J., Benkö, P., Snoeren, A. C., Voelker, G. M., & Savage, S. (2006). *Jigsaw: solving the puzzle of enterprise 802.11 analysis* (Vol. 36, No. 4, pp. 39-50). ACM.

[169] Gupta, A., Min, J., & Rhee, I. (2012). WiFox: Scaling WiFi performance for large audience environments. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies* (pp. 217-228). ACM.

[170] Raya, M., Aad, I., Hubaux, J. P., & El Fawal, A. (2006). DOMINO: Detecting MAC layer greedy behavior in IEEE 802.11 hotspots. *Mobile Computing, IEEE Transactions on*, *5*(12), 1691-1705.

[171] Qiu, L., Bahl, P., Rao, A., & Zhou, L. (2005). Troubleshooting multihop wireless networks. In *ACM SIGMETRICS Performance Evaluation Review* (Vol. 33, No. 1, pp. 380-381). ACM.

[172] Yan, B., & Chen, G. (2009). Model-based fault diagnosis for IEEE 802.11 wireless LANs. In *Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous, 2009. MobiQuitous' 09. 6th Annual International* (pp. 1-10). IEEE.

[173] Sheng, Y., Chen, G., Yin, H., Tan, K., Deshpande, U., Vance, B., & Wright, J. (2008). Map: a scalable monitoring system for dependable 802.11 wireless networks. *IEEE Wireless Commun.*, *15*(5), 10-18.

[174] Yeo, J., Youssef, M., & Agrawala, A. (2004). A framework for wireless LAN monitoring and its applications. In *Proceedings of the 3rd ACM workshop on Wireless security* (pp. 70-79). ACM.

[175] Münz, G., Li, S., & Carle, G. (2007). Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet*.

[176] Wireless Infrastructure Management, http://www.ca.com/ last accessed November 2014.

[177] Wireless Security Auditor, http://www.research.ibm.com/gsal/wsa/ last accessed November 2014.

[178] AirMagnet. AirMagnet Distributed System. http://airmagnet.com/ last accessed November 2014.

[179] AirDefense. Wireless LAN Security. http://airdefense.net/ last accessed December 10, 2014

[180] Rojas, A., Branch, P., & Armitage, G. (2005). Experimental validation of the random waypoint mobility model through a real world mobility trace for large geographical areas. In *Proceedings of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems* (pp. 174-177). ACM.

[181] Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender Systems Handbook* (pp. 39-71).

[182] Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, *148*(3), 929-936.

[183] Posada, D., & Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics,* 14(9), 817-818.

[184] Hillmer, S. C., & Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association,* 77(377), 63-70.

[185] Mc Clary, D. W., Syrotiuk, V. R., & Kulahci, M. (2010). Profile-driven regression for modeling and runtime optimization of mobile networks. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *20*(3), 17.

[186] Chan, A., Zeng, K., Mohapatra, P., Lee, S. J., & Banerjee, S. (2010). Metrics for evaluating video streaming quality in lossy IEEE 802.11 wireless networks. In *INFOCOM, Proceedings IEEE* (pp. 1-9). IEEE.

[187] Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691-692.

[188] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In Kdd Vol.* 96, pp. 226-231.

[189] Event Processing with Esper, http://esper.codehaus.org/ last accessed November 2014.

[190] Turner, R., Ghahramani, Z., & Bottone, S. (2010). Fast online anomaly detection using scan statistics. In *Machine Learning for Signal Processing (MLSP), IEEE International Workshop on* (pp. 385-390). IEEE.

[191] Arasu, A., & Widom, J. (2004). Resource sharing in continuous sliding-window aggregates. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30* (pp. 336-347). VLDB Endowment.

[192] Dujovne, D., Turletti, T., & Filali, F. (2010). A taxonomy of IEEE 802.11 wireless parameters and open source measurement tools. *Communications Surveys & Tutorials, IEEE*, *12*(2), 249-262.