ON THE DESIGN OF UNIVERSAL SCHEMES FOR MASSIVE

UNCOORDINATED MULTIPLE ACCESS

A Thesis

by

AUSTIN AHMAD TAGHAVI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Jean-Francois Chamberland-Tremblay
Co-Chair of Committee,	Krishna Narayanan
Committee Members,	Andrew Jiang
	Nicholas Duffield
Head of Department,	Miroslav Begovic

May 2017

Major Subject: Computer Engineering

ABSTRACT

The wireless landscape is shifting to include more networks with massive numbers of unattended devices, seeking to distribute their information sporadically. This paradigm shift has spawned a lot of interest in the adaptation of multiple access schemes to accommodate such systems. In particular, slotted ALOHA based systems have been largely considered. Researchers have noted the analogous relationship between erasure coding theory and slotted ALOHA based multiple access schemes. This connection has allowed researchers to leverage erasure coding theory research, and use this research to improve slotted ALOHA based multiple access schemes using successive interference cancellation. In this thesis, we carry on with this line of research, and extend the work to consider slotted ALOHA based schemes with the constraint that the number of users in the system is not known. This constraint is not only novel, but also widely applicable to the modern wireless landscape. In particular, systems that are part of the Internet of Things (IoT) may necessitate systems that perform well even when the number of users in the system is unknown. We propose a transmission strategy for active devices based on Markov chains. In addition, we derive necessary and sufficient conditions for probability distributions to be shaped by such Markov chains. Numerical results show that, even with this constraint, a significant improvement to the performance of slotted ALOHA is attainable. In addition, we seek to explore other novel formulations of the uncoordinated slotted multiple access problem that also do not have knowledge of the number of users in the system, but include multiple access points with overlapping users. For this problem formulation, we show that a shared decoding process between the access points provides substantial performance improvements.

ACKNOWLEDGEMENT

I would like to thank my committee chair and advisor, Dr. Jean-Francois Chamberland-Tremblay for his constant guidance and support during the course of this research. In addition, I would like to thank Dr. Krishna Narayanan and Avinash Vem for their collaboration during this research. Finally, I would like to thank my fellow graduate students, family, and friends for making my graduate education such an enjoyable and fulfilling experience.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professors Jean-Francois Chamberland-Tremblay (advisor), Krishna Narayanan, and Nicholas Duffield of the Department of Electrical Engineering and Professor Anxiao Jiang of the Department of Computer Science.

The analysis in Chapter II, Chapter III, and Chapter IV was done in collaboration with Dr. Jean-Francois Chamberland-Tremblay, Dr. Krishna Narayanan, and Avinash Vem of the Department of Electrical Engineering, and published in 2016. All other work conducted for the thesis was completed by the student independently, with the advising of Dr. Jean-Francois Chamberland-Tremblay.

Funding Sources

TABLE OF CONTENTS

Page

LIST OF FIGURES

1. INTRODUCTION*

With the emergence of machine-driven wireless communication, we are witnessing a change in the wireless landscape. The traditional traffic generated by individuals and their usage of personal phones and mobile computers is being supplemented with traffic generated by a massive amount of unattended devices that seek to communicate their messages sporadically. The Internet of Things (IoT) represents a large number of such devices that break the mold of traditional communication paradigms. This paradigm shift is not trivial; it drastically changes the profile of a typical subscriber, which impacts the performance of currently deployed protocols. Thus, this landscape change necessitates the design of new random access schemes that are better suited to such traffic. Specifically, there is a need for the design of random access schemes where we expect a massive number of users to act in a largely uncoordinated fashion. In this research, we consider multiple novel formulations of random access problems, and explore the design space of transmission protocols tailored to this evolving wireless landscape.

To discuss random access schemes, we must first consider the flagship protocol: ALOHA (Additive Links On-line Hawaii Area). Developed by Norman Abramson, the ALOHA framework, also known as ALOHAnet, is a system for wireless and wired computer communication. One of the original goals of the ALOHA project was to allow communication between many devices and an access point without using a unique frequency for each device. The spirit of this protocol is to let users communicate their information with the access point in an uncoordinated fashion.

The protocol uses two frequencies in a hub/star configuration. One frequency is used for the access point to send messages to the connected devices, and the other is used for the devices to transmit messages to the access point. The protocol itself does not tell devices when to transmit their data. In the event that two or more devices transmit their data at the same time, the access point sends a message to each of those devices letting them know that a collision has occurred and that they should retransmit their messages with a random delay. An improvement on ALOHA, which divided transmission times into discrete time intervals, became known as slotted ALOHA. The intuition behind slotted ALOHA is to minimize collisions by only allowing devices to transmit their data at the beginning of a "time-slot" rather than simply allowing devices to transmit whenever they have data to transmit. This improvement to the ALOHA framework does require a non-trivial addition: global time synchronization. All devices must be capable of determining the exact start of a time-slot. While this adds a measure of complexity, the performance advantages are well worth the extra effort in many fields. In general, most of the following research we will discuss is related to the slotted ALOHA framework rather than the original version, often called "pure ALOHA". The core intuition behind ALOHAnet is largely referred to as the concept of random access. Random access still plays an important role in several wireless communication today. For example, slotted ALOHA is largely used in areas such as subscriber-based satellite communications networks and RFID tag identification.

There are some general conditions and framework pre-requisites for ALOHA and derivative protocols that are assumed. For instance, there are always assumed to be a fairly large number of devices of users in the system. In addition, all of these devices communicate over a shared channel with an access point. The performance of these systems is often measured by the throughput T, most often defined as the

average number of packets decoded per time slot, in the case of slotted ALOHA derivative protocols. In general, there is no communication assumed between the devices themselves.

As mentioned above, some of the primary areas of communication that use ALOHA and slotted ALOHA based protocols are radio-frequency identification (RFID) and satellite communications. In the RFID field, the process for which ALOHA based protocols are used is called singulation. Singulation is the act of an RFID reader singling out a specific tag from a field of tags. An example is that when scanning a bag of groceries, a reader will need to single out each individual item and try to avoid collisions. Because these tags transmit on the same frequency, it is clear that ALOHA and ALOHA based protocols are good candidates for this type of communication. Satellite communication was one of the original uses of the ALOHA protocol, and ALOHA-based protocols are still largely used in satellite communications. The use of a shared communication medium for each user in the network is what makes ALOHA based protocols ideal for satellite communications, where propagation time is lengthy and high levels of coordination are often not possible.

Of course, ALOHA and slotted ALOHA have some considerable drawbacks. While revolutionary and very computationally manageable, the protocol's performance leaves something to be desired in certain scenarios. For example, the (access) throughput for slotted ALOHA is limited by $1/e$. This precludes the protocol from being used in certain situations where a higher throughput is necessary. Because of this, over the past few decades, much attention in the research community has been focused on improving ALOHA and specifically slotted ALOHA to improve the throughput limit. It is important to note that other protocols, such as demand assignment multiple access and medium access control, have been introduced and serve a somewhat similar purpose with good performance. However, in many situations,

such as satellite communications, slotted ALOHA is still a very appealing paradigm, in particular because of the low latency access associated with the protocol. Because of the inherent advantages of random access, much research attention has been devoted to random access protocols and improvements as well as theoretical limits for various scenarios. Variations and improvements upon the ALOHA protocol have become and still are standards for communication in many industries.

An early, but very important, variation of the ALOHA protocol was introduced in the paper "Diversity ALOHA-A Random Access Scheme for Satellite Communications", by Choudhury and Rappaport, in 1983 [3]. The key novelty of this research is that each device transmits multiple copies of their message in the first place. This can be thought of as a generalization of the ALOHA platform. Two methods are proposed in the paper: frequency diversity and time diversity. In frequency diversity, the same message is transmitted at the same time, but on different frequencies. In time diversity, the same frequency is used, but a random time delay is employed in between each transmission. Time diversity can be thought of as simply acting as if a collision has occurred before being notified by the access point in traditional slotted ALOHA. In traditional ALOHA, a device must wait a propagation time T for the access point to indicate that a re-transmission is necessary. In practice, the packet duration τ is must shorter than T. The time diversity proposed in this research suggests that, rather than wait the propagation time T to decide whether or not to re-transmit the message, a device should simply re-transmit k times with random delays. This scheme is shown to have a much lower expected delay for a given desired throughput, given that the throughput is a certain degree lower than the maximum. In cases where the highest throughput is desired, the re-transmission proves wasteful compared to the traditional scheme. The idea of time diversity has serves as a foundation for many other more modern variations of the ALOHA protocol.

Of particular relevance to our research is a modification to slotted ALOHA referred to as Contention Resolution Diversity Slotted ALOHA (CRDSA) [2] This scheme uses diversity transmission of data bursts and successive interference cancellation techniques to improve upon the performance of traditional slotted ALOHA. The primary contribution of the CRDSA scheme is that it proposes a method of using a successfully transmitted packet to cancel the interference of that packet with other packets in different time slots. This process is repeated iteratively to decode as many packets as possible. This scheme provides a peak throughput of 0.55, a significant improvement over traditional slotted ALOHA. In particular, the CRDSA scheme provides a 17 fold throughput increase over traditional slotted ALOHA and a 4.5 fold throughput increase over diversity slotted ALOHA. The probability of collisions in subsequent bursts and in preamble transmissions are derived analytically as well. The research in [2] has spawned an entire category of slotted ALOHA-based protocols in the modern era of random access. An extended version of CRDSA, called CRDSA++ by the researchers, is proposed in [5]. This protocol considers more repetitions than CRDSA originally did, and also attempts to use the information of the power in the received signal to further increase performance. A packet loss ration of less than 10^{-4} is shown to be possible, an improvement over CRDSA. Liva et al. [9] further this line of thinking and provide significant performance improvements over CRDSA. Liva et al. make the connection between the interference cancellation process of CRDSA and a bipartite graph. This leads to a connection between the CRDSA scheme and the iterative erasure coding of graph based codes. Liva et al. show that by leveraging the research done on iterative erasure coding of graph based codes to select particular distributions for the nodes of the graph representing the random access transmissions, a very high throughput rate can be achieved. A throughput of $T = 0.97$ can be achieved for large frames and $T = 0.8$

can be achieved for practical implementations. Finally, and very relevant to the research contained herein, is work done by Narayanan and Pfister [13]. This work follows in the same line of research as Liva et al. [9], and shows analytically that the Soliton distribution is optimal for the distribution of the nodes in the bipartite graph representing the random access transmissions. In addition, this work shows that the throughput of such a transmission scheme can be arbitrarily close to one. The Soliton distribution is a distribution proposed by Luby [10] as an ideal distribution for symbols in what he describes as "universal erasure codes". However, the work by Liva, et al. connecting the random access schemes to erasure code theory provides the foundation for a link between the Soliton distribution and slotted ALOHA.

It is worth giving a brief overview of erasure codes, to give a better understanding of their application to slotted ALOHA-based communication protocols. Erasure codes are a class of error correcting codes. An erasure code takes a message of length k symbols and transforms it into a longer encoded message of n symbols. The goal is for the original message to be recoverable from a subset of the n symbols. The code rate is defined as the length of the original sequence of symbols divided by the length of the encoded sequence of symbols, mathematically k/n. The relation to slotted ALOHA is fairly intuitive. Each device and its message is analogous to a single symbol in an erasure code, and the n symbols are analogous to the n time slots in a slotted ALOHA protocol. The kind of erasure codes that are directly applicable to slotted ALOHA are fountain codes [11]. Fountain codes are a type of near-optimal erasure codes, meaning that they require $(1 + \epsilon)k$ encoding symbols to ensure decoding of a message of length k symbols. The analogy given in the original paper introducing fountain codes [11], describes the intuition for fountain codes as follows: An original message consisting of Kl bits is to be encoded. The encoder is a fountain that drips encoded packets of length l. If one wishes to decode the

6

original message, they can hold a bucket under the fountain until they have collected a little more than K drops. At this point, the message can be decoded. Fountain codes are considered to be rateless, meaning that there are an unending number of encoded symbols that can be generated from any given original sequence of symbols. A particular novelty of these codes is that the number of encoded symbols needed does not need to be determined a priori. The codes are considered universal because they are near-optimal for any erasure channel; no matter the erasure probabilities of the channel, the encoded can keep sending symbols until the original message can be decoded. It is clear that these codes are particularly applicable to slotted ALOHA based schemes. The access point can simply keep the transmission round going until it has enough time slots to decode the message for each device (or a desirable portion of the devices). Because of the progress made with CRDSA in using successive interference cancellation to decode messages for slotted ALOHA based schemes, the digital fountain abstraction can be applied to such schemes. Although the idea of fountain codes and universal rateless erasure codes was introduced in [11], the first actual realization of such codes came from Luby in [10]. He refers to these codes as Luby Transform, or LT codes. As with fountain codes, LT codes ensure that the original sequence of k symbols can be recovered from any set of encoded symbols that is slightly larger than k. The process for generating an encoding symbol for an LT codes is fairly simple. First, the degree d of the encoding symbol is chosen from a degree distribution. Next, d of the original k input symbols are chosen at random, uniformly. Finally, the encoded symbols is simply the exclusive-or of the chosen d input symbols. The primary focus of the research for LT codes is in choosing optimal distributions for choosing the encoding symbol degree d. The above mentioned Soliton distribution is the distribution that is suggested to be optimal by Luby in [10]. The relation of these erasure codes to the structure of communication

for slotted ALOHA-based schemes has made an entire branch of research available to be applied to random access protocols. In particular, our contributions follow the line of research that is exploring the application of erasure coding theory to random access schemes.

Ivanov et al. explore in [6] the error floor of coded slotted ALOHA-based protocols. Much of the research effort into slotted ALOHA based protocols are focused mainly on maximizing the throughput of a protocol. However, there certainly exist scenarios in which the reliability of the communication is worth optimizing at the expense of some amount of throughput. Ivanov et al. in [6] derive analytical expressions for the error floor of coded slotted ALOHA protocols. These findings are valuable to our research effort because it provides analytical results that are relevant to the scenario in which percentage of nodes decoded is more important than solely the throughput.

These results have improved upon slotted ALOHA to achieve throughput levels that are high both in the limit and in practical implementations. Yet, these results largely rely on the system having knowledge of the number of users in the system. In light of the changing wireless landscape, and with the knowledge that the trend of largely unattended devices is likely to continue, it is reasonable to assume that the number of users in such a system may not be known, and it would be wise to design the transmission protocols to perform well in such circumstances. It is not difficult to imagine models where the above discussed transmission schemes based on slotted ALOHA will not perform as well or will not be applicable. It is worth mentioning that some research has been focused on addressing this issue. Stefanovic et al. [14] propose a scheme in which the number of users is estimated at the end of a transmission round. The primary novelty of this protocol is that the duration of transmission rounds is not decided on before the initial transmission round. Instead, a transmission round

continues until an acceptable estimation can be made and message resolution is satisfactory. An example of research into the modern implementation of an advanced slotted ALOHA based protocol is presented in [8], where Lee et al. propose a slotted ALOHA-based protocol for RFID tag identification. Lee et al. note that slotted ALOHA can become rather inefficient in RFID identification systems, as the number of time slots needed to recognize all of the tags can grow rapidly as the number of users in the system increases. The proposed solution to this issue is a protocol called Enhanced Dynamic Framed Slotted ALOHA. The key novelty to this protocol is that, at the end of a transmission round, the access point estimates the number of unread RFID tags and uses this information to make the proper changes to the number of responding tags and the number of time slots in a transmission round. The purpose of this strategy is to increase the efficiency of the system by performing smaller transmission rounds for the users that were not decoded until the desired number of users have their messages decoded. Their results show an $85 - 100\%$ increase in the efficiency of this protocol compared to more conventional protocols. Of course, the drawback of this research if the increased level of coordination in between transmission rounds.

The idea of a random access scheme based on slotted ALOHA that is agnostic of the number of users in the network and does not make an attempt to estimate the number of users is a largely unexplored research topic. It is a relevant topic because a model that requires or desires such a transmission scheme is very feasible in today's wireless landscape, and will only become more common in the future. In particular, the aforementioned scheme involving estimating the number of users in the network [14] may leave some throughput performance to be desired in cases where the number of users changes sporadically and is not easily estimated. We seek to build upon the work done by Liva et al. [9] and Narayanan, Pfister [13] and attempt to apply

9

their results to a formulation of the slotted multiple access problem in which the access point does not know and does not attempt to estimate the number of users in the network. In this research, we consider a single round of transmission with the following question in mind: Is it possible to design a universal multiple access scheme and transmission policy for the active devices without having knowledge of the number of users in the system? We can identify three primary contributions of this facet of the research. First, we propose a slotted ALOHA problem formulation that is both novel and practical. In this formulation, the access point does not know the number of active devices in the system, nor does the access point need to estimate the number of active users. In addition, the number of time slots within a transmission round is not known a priori. This formulation of the slotted multiple access problem is both relevant to the modern wireless landscape, and unexplored. Second, this research proposes a Markov framework for transmission slot selection by active devices. This Markov strategy provides each device with a method for deciding whether or not to transmit at each time slot based on two parameters: the current time slot in the transmission round, and the number of times the devices has already transmitted its message during the current transmission round. Note that this strategy is independent of other devices, the total number of time slots, and the total number of active devices. This strategy could be deployed and communicated to devices with minimal coordination, making it ideal for the slotted multiple access formulation that we are interested in. Lastly, we show conditions that a sequence of probability distributions must meet to be shaped by such a Markov strategy. In addition, we show that reasonable distributions are achievable with the Markov strategy. In particular, we show that the Soliton distribution, which has been shown to be an ideal distribution for the slotted multiple access with successive interference cancellation, is achievable through the Markov framework.

An ongoing leg of this research is the exploration of other novel formulations of the slotted multiple access problem. More specifically, we consider the reality that many of the scenarios that involve an implementation of a random access scheme involve not one, but multiple access points. As such, we believe that there is something to be gained from considering these models and the interactions of multiple access points with possible overlapping devices, rather than only considering each access point in isolation. We consider a formulation of the slotted multiple access problem similar to the one used in the other areas of this research. In this formulation there are two access points, and a portion of the active devices transmit to both access points, while the other devices transmit to only of the two access points. While the single access point formulation can be represented with a single bipartite graph, as shown in [9], this latter formulation can be represented as two connected bipartite graphs. In addition, the iterative successive interference cancellation process can be modified and extended to work with these connected bipartite graphs.

The concept of using multiple access points for an ALOHA based protocol was introduced by Corson and Ephremides [4]. This research was motivated by the need for more efficient communication for mobile users in factories. The underlying idea is to use multiple radio base stations as receivers to implement a multi-receiver slotted ALOHA protocol. This research explores the idea of having multiple receivers with the same footprint, and also with separate but possibly overlapping footprints. The throughput analysis of this model shows that there is a significant throughput decrease with large overlapping areas for the receivers, and suggests that minimizing overlap is ideal for throughput. Here, throughput is being defined in terms of total number of messages decoded per time slot, considering the total number of time slots to be nk, where n represents the number of time slots per receiver and k denotes the number of receivers.

In [12], Liva et al. continue to explore the idea of multiple receivers with the same footprints, dubbing the concept "spatial diversity". In the model discussed in [12], there exist K access points. Each active device in the network transmits messages in the same manner as other slotted ALOHA based systems. The key novelty in this model is that the communication channel between the active devices and each of the K access points is considered to be an erasure channel with erasure probability ϵ. However, this effort is largely concerned with standard implementations of slotted ALOHA, and their model differs from the model we wish to explore. Liva et al. derive theoretical throughput limits and efficiency ceilings of slotted ALOHA schemes with spatial diversity, as well as optimal values for ϵ for the erasure channels. In addition, a model for communication between the access points is derived, and optimal strategies for this communication are obtained. A small portion of this research is devoted to introducing the idea of using successive interference cancellation in conjunction with their spatial diversity model. A significant throughput performance increase is noted when successive interference cancellation is used, and the tradeoffs between the performance increase and the added complexity of successive interference cancellation are discussed. In [7], Jakovetic et al. explore the same model, but further the effort to find optimal strategies for slotted ALOHA with spatial diversity using successive interference cancellation for decoding. This work compares three decoding processes for throughput and efficiency. The models are spatial cooperation, temporal cooperation, and spatial-temporal cooperation. The spatial-temporal cooperation model is further explored, and optimal degree distributions are derived for this model.

Zorzi et al., in [15], explores the performance of slotted ALOHA systems with spatial diversity in the presence of Rayleigh fading. The key idea behind this research is to use the spatial diversity (having multiple receivers) and have the access point

combine the received signals from multiple receivers in a meaningful way. A significant performance gain of 33% is achieved with using two receivers in the presence of Rayleigh fading when compared with the conventional setting. Of course, adding this spatial diversity introduces complexity that is not present in single receiver systems.

For mobile networks, a related line of research is that of Coordinated Multipoint (CoMP) systems. Coordinated Multipoint is a set of techniques for LTE communication, which often employ ALOHA based protocols. In these systems, the data that is transmitted by each device can be received by multiple receivers, the receivers being eNodeB's. Using the resource allocation techniques that are implemented in LTE, the multiple receivers are used to be able to receive a message from a user even when that message would traditionally be lost due to poor conditions such as shadowing. One suggested solution is essentially a MIMO system created by the eNodeB's.

While the above works explore using multiple access points with slotted ALOHA and successive interference cancellation, their model differs in a key way from the model we wish to explore. The primary difference is that our model considers multiple access points with overlapping footprints, and that active devices that fall in the overlapping area will always transmit to both (or all) access points. The previous research in [15] and [7] considers a model with multiple access points, but that every device can transmit to all access points. This model uses an erasure channel for the link between each device and each access point as a way of creating diversity in transmissions. Other models consider multiple access points and their effect on systems with Rayleigh fading and shadowing.

Herein, we consider the multiple access point random access problem formulation described above. First, we consider the resulting bipartite graph structure that is induced by this problem formulation. We also consider the relevant decoding process, and show results for the performance increase of sharing information between access

points during the decoding process using successive interference cancellation. Finally, we explore the structure of these graphs and which distributions are ideally suited for them, and continue this line of thinking to the Markov framework described above.

2.   PROBLEM FORMULATION*

2.1   System Model

As mentioned, the problem we are interested in is random multiple access over wireless infrastructures. This problem was first explored in [1]. Since then, it has become a rather large and diverse area of research, and implementations of such protocols are widely used in certain forms of communication. Figure 2.1 is a high level diagram that depicts the primary components of slotted multiple access, as discussed in this thesis. Our problem formulation is closely related to the problem formulation in content resolution diversity slotted ALOHA (CRDSA) [2]. We are concerned with the uplink communication: the messages that are sent from the active devices in the system to the access point. As in slotted ALOHA, the possible transmission times are discretized into time slots. Any message that is scheduled will be transmitted to the access point at the start of the next time slot. At the onset of every transmission round, the access point notifies the active devices that a transmission round is starting, which signals them to begin transmitting their messages. Each device has a single message to transmit over the transmission round, and repeatedly transmits that message during a random subset of the time slots during the round. It should be noted that neither the access point nor the active devices know the length of the transmission round a priori. When a message is received at the access point without interference, that is, when a message is the only message transmitted during a given time slot, it is assumed to be decodable. When

---
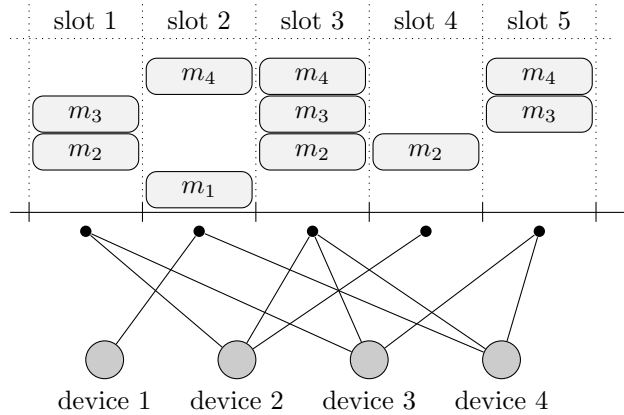
Figure 2.1: Bipartite graph representing a transmission round for the described random access scheme.

two or more devices transmit their message during the same time slot, there is a collision at the access point during that time slot. We also assume that once a given message is decoded, it contains information that allows the access point to determine all of the other time slots during the current transmission round when this message was transmitted. The access point then uses successive interference cancellation to attempt to iteratively decode as many of the messages as possible. This continues until the end of the transmission round, as decided upon by the access point. The end of the transmission round is signaled to the active devices, along with information about which packets were successfully decoded, if desired.

Content resolution diversity slotted ALOHA has been well studied. Because of its structural relation to erasure codes [9], approaches from the field of erasure codes have been applied to CRDSA protocols. Optimal distributions and theoretical throughput limits have been analytically derived [13]. Still, this problem has not been as deeply explored from the angle of having an unknown number of users. In [14] , the number of users is estimated at a certain point and used as a truth thereafter. In this research, we explore the CRDSA problem formulation where the number of

16

users is not known and is explicitly not estimated. Neither the access point nor the active devices have knowledge of the number of active users, and the number of time slots in a transmission round is not decided a priori. It is worth noting that we are not focusing on downlink traffic or implementation details. Rather, we are solely focused on uplink traffic and transmission schemes for devices to employ.

2.2   Performance Metrics and Notation

In this section, we define pertinent notation and performance metrics used throughout this thesis. For a given transmission round, we will refer to N as the number of time slots in the transmission round. We will refer to the number of active devices transmitting messages during a given round as K. For example, in Figure 2.1, we see a graphical representation of a transmission round. In this round, there are $N = 5$ time slots and $K = 4$ active devices transmitting messages. Worth noting will be the degree of each node in the graph. In terms of the bipartite graph structure, we will refer to the nodes that represent an active devices as "variable nodes". Similarly, we will denote the nodes that represent time slots as "check nodes". In a given bipartite graph representing a single transmission round, the degree of a variable node will be equal to the number of times that the active device represented by that variable node transmitted its message during that transmission round. Similarly, the degree of a given check node will be equal to the number of messages that were transmitted to the access point during the time slot that is represented by that check node. We will refer to the number of decoded messages at the end of a transmission round as D.

For performance assessment, the primary metric that we use is the throughput of a protocol. This is defined as the number of messages successfully received and decoded per time slot. Equivalently, it is defined as $T = D/N$. The performance

17

users is not known and is explicitly not estimated. Neither the access point nor the active devices have knowledge of the number of active users, and the number of time slots in a transmission round is not decided a priori. It is worth noting that we are not focusing on downlink traffic or implementation details. Rather, we are solely focused on uplink traffic and transmission schemes for devices to employ.

2.2   Performance Metrics and Notation

In this section, we define pertinent notation and performance metrics used throughout this thesis. For a given transmission round, we will refer to N as the number of time slots in the transmission round. We will refer to the number of active devices transmitting messages during a given round as K. For example, in Figure 2.1, we see a graphical representation of a transmission round. In this round, there are $N = 5$ time slots and $K = 4$ active devices transmitting messages. Worth noting will be the degree of each node in the graph. In terms of the bipartite graph structure, we will refer to the nodes that represent an active devices as "variable nodes". Similarly, we will denote the nodes that represent time slots as "check nodes". In a given bipartite graph representing a single transmission round, the degree of a variable node will be equal to the number of times that the active device represented by that variable node transmitted its message during that transmission round. Similarly, the degree of a given check node will be equal to the number of messages that were transmitted to the access point during the time slot that is represented by that check node. We will refer to the number of decoded messages at the end of a transmission round as D.

For performance assessment, the primary metric that we use is the throughput of a protocol. This is defined as the number of messages successfully received and decoded per time slot. Equivalently, it is defined as $T = D/N$. The performance

of a protocol can be measured by the expected throughput of a transmission round using that protocol. We will also consider the limit of the throughput as the number of devices goes to infinity. Often, when measuring the performance of a transmission scheme, the ratio of devices to time slots will be fixed, and the throughput will be measured for many different values for the number of active devices.

2.3 Decoding Strategy

As discussed in [2], the process of successive interference cancellation can be applied to slotted ALOHA based schemes to attempt to successfully decode packets even in time slots where collisions occurred. In implementation, this means that every transmitted message needs to have some metadata about the other time slots during which the message was transmitted during that transmission round. The access point can then use interference cancellation to remove that message from any collisions that it was involved in. The goal is to decode as many messages as possible with this process. The iterative decoding process of decoding is easily defined in terms of the bipartite graph as follows:

Algorithm 1 Successive Interference cancellation.
---
  procedure Peeling Decoder
      `Degree1Nodes` ← `Nodes` of degree one
      while `Degree1Nodes.size()` > 0 **do**
         for `Node` **in** `degree1nodes` **do**
            Decode `Node`
            Cancel interference produced by `Node`
            `DecodedNodes.append(Node)`
            Remove `Node` and `Node.edges()` from graph
            Remove `Node` from `Degree1Nodes` and `Nodes`
         `Degree1Nodes` ← `Nodes` of degree one

2.4   Node Distributions

In a typical slotted ALOHA based protocol without the use of successive interference cancellation for message decoding, the criterion for maximizing the throughput is clear. The only decodable messages are those that were the only messages transmitted during a given time slot. Thus, the expected throughput for a given number of time slots, N, and a given number of devices K, will be directly proportional to the number of degree one check nodes, given uniformity of the transmission policy for all devices. However, when using successive interference cancellation to decode messages at the end of a transmission round, it is less clear how to optimize the throughput in terms of the degrees of the nodes. Yet, because of the relation to erasure codes, there is a field of research that can be leveraged to find optimal node distributions for decoding with successive interference cancellation. LT codes offer valuable insight in picking degree distribution for slotted ALOHA with successive interference cancellation. The research done in [13] analytically derives the optimal distributions for the slotted ALOHA with successive interference cancellation. The researchers found the Soliton distribution to be asymptotically optimal. Indeed, the Soliton distribution was initially proposed in [10] as a suitable distribution for choosing the degree of encoding symbols for LT codes, yet it also applies to the dual construction that arise in uncoordinated multiple access. For $t \in \mathbb{N}$, the probability distribution function for the Soliton distribution is as follows:

$$
p_{\mathrm{sol}(t)}(m) = \begin{cases} \frac{1}{t} & m = 1 \\[2ex] \frac{1}{(m-1)m} & m = 2, \dots t. \end{cases}
$$

We note that, in the context of uncoordinated multiple access, t in the above definition represents the number of time slots and m denotes the degree of a specific

19

variable node. The implementation for a slotted ALOHA scheme can be imagined as follows. At the beginning of each transmission round, the access point decides how many time slots the upcoming round will consist of, and communicates this to the active devices. Each active device employs this information to decide how many times to transmit their message during the transmission round, drawing from the Soliton distribution. Each device chooses, uniformly, a random subset of time slots with cardinality equal to the number of times they wish to transmit their message. At the end of the transmission round, the access point decodes as many messages as possible using successive interference cancellation, as detailed in Algorithm 1.

This approach cannot be directly taken in the scenario that we have proposed. In our problem formulation, we have specifically stated that neither the access point nor the devices themselves know the number of users or attempt to estimate the number of users, and that the number of time slots in a transmission round is also not known a priori. Because of these two conditions, the access point cannot transmit the number of time slots for the upcoming transmission round to the users. This means that the users cannot use a distribution that is dependent on the total number of time slot to decide how many times to transmit their message or during which time slots to transmit. Any such distributions with reasonable performance depends on the total number of time slots in the transmission round as a parameter to the distribution, so the problem formulation we have described necessitates a different abstract framework.

3. MARKOV FRAMEWORK*

Our problem formulation cannot use the more traditional approach of having active devices simply draw their degree from a distribution and then randomly choose during which time slots to transmit their message. Our solution is to use a Markov framework to achieve the desired degree distributions for the variable nodes. More specifically, we propose a transmission policy that is implemented as a Markov chain, for each active device. The transmission policy tells the active device, at the onset of each time slot in a transmission round, whether or not to transmit its message. This transmission policy will not need any knowledge of the total number of users in the system or the total number of time slots in a transition round. We also refer to such transmission policies as "uncoordinated transmission policies". In this thesis, we solely consider the scenario in which all active devices use the same policy. The motivation for using the Markov framework is that devices can simply follow a transmission policy until the access point decides that the transmission round is complete.

3.1 Abstract Framework

We define our uncoordinated transmission policy for a given device as follows. The policy takes in the current time slot t, and the number of times that the device has already transmitted its message during the current transmission round. In response to the given input, the policy gives a 1 or a 0 as output. A one indicates that the device should transmit its message, whereas a zero indicates that the device should not transmit its message. We denote a policy by $\boldsymbol{\mu} = (\mu_0(\cdot), \mu_1(\cdot), \ldots)$, where $\mu_t(m)$

---

*Part of the work in this section is reprinted with permission from "On the design of universal schemes for massive uncoordinated multiple access" by Austin Taghavi, Avinash Vem, Jean-Francois Chamberland, and Krishna Narayanan. 2016. ISIT July 2016: 345-349 by IEEE. ©2016 IEEE.

represents the (random) decision rule for time slot t with the number of messages sent m as a parameter. Using such a policy, the number of times a packet is transmitted during a transmission round by a specific device can be represented a random process X_t. The random process X_t is defined recursively as

$$X_{t+1} = X_t + \mu_t(X_t) \quad t \geq 0. \tag{3.1}$$

For a given policy $\boldsymbol{\mu}$, we denote the distribution of the resulting Markov process X_t at the time of time slot t as p_t. It is worth noting that all devices employ the same policy, and that the policies (and thus the degree of all variable nodes) are independent. Because of the recursive nature of (3.1), a natural interpretation for X_t is a Markov chain. In this Markov chain, the state of the chain for a given device is equivalent to the number of times that the device has transmitted its message during the current transmission round. Thus, X_0 is necessarily 0. In such a Markov chain, at each time slot, the chain can either move up one state or not at all. This is because a device can transmit at most one message per time slot, so the total number of messages sent can only increase by one at each time slot. The transition probabilities for the Markov chain can be expressed in terms of X_t and $\mu_t(m)$ as follows:

$$\Pr(X_{t+1} = m | X_t = m) = \Pr(\mu_t(m) = 0)$$

$$\Pr(X_{t+1} = m + 1 | X_t = m) = \Pr(\mu_t(m) = 1).$$

Furthermore, we can infer that a message is sent at time slot t whenever $X_t - X_{t-1} = 1$. In terms of the Markov chain, the earlier described p_t can also be defined as the probability distribution of the Markov chain X_t at time slot t. Finally, we note that
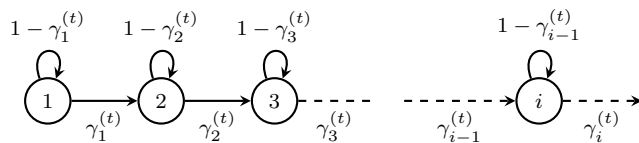
Figure 3.1: Example of a Markov chain interpretation of a simple transmission policy.

the unconditioned probability that a given device will transmit its message during time slot t can be calculated as the Wasserstein distance between the distributions p_t and p_{t+1}:

$$\Pr(X_{t+1} = X_t + 1) = W_1(p_t, p_{t+1}) = \inf_{\gamma \in \Gamma(p_t, p_{t+1})} \sum_{\mathbb{N}_0 \times \mathbb{N}_0} \|m - n\|_1 d\gamma(m, n)$$

where $\Gamma(p_t, p_{t+1})$ represents the collection of joint measures on $\mathbb{N}_0 \times \mathbb{N}_0$ with marginals $p_t(\cdot)$ and $p_{t+1}(\cdot)$ on the first and second factors, respectively.

3.2 Distribution Shaping

In this section, we focus on creating Markov chains for certain distributions for the degree of variable nodes. To do this, we must first define necessary and sufficient conditions that a distribution must meet to be shaped into a transmission policy Markov chain. The conditions that determine if a sequence of probability distributions can be shaped are largely based on stochastic dominance. We consider X and Y to be random variables drawn according to probability distributions p_X and p_Y, respectively. In this paper we will write $X \preceq Y$ and $p_X \preceq p_Y$ to denote first-order stochastic dominance. This is defined as $X \preceq Y$ when $\Pr(X > \ell) \leq \Pr(Y > \ell)$ for all $\ell \in \mathbb{R}$. In particular, the Markov chains we are interested in are monotone increasing, only over the set of non-negative integers $\mathbb{N}_0$, and time inhomogeneous. In addition, these Markov chains only have self-transitions, and transitions to the

closest neighbor to the right.

Proposition 1. *Let $(p_0, p_1, ...)$ be a sequence of probability distributions over $\mathbb{N}_0$. This sequence of distributions is achievable through the above described class of Markov chains if and only if*

$$p_t \preceq p_{t+1} \quad and \quad p_{t+1} \preceq Sp_t \quad \forall t \in \mathbb{N}_0 \tag{3.2}$$

where S is the standard right shift operator as applied to one-sided infinite sequences of numbers. The right shift operator is simply a translation of a probability distribution by shifting the probabilities to the right by one. If $p_t = (a_1, a_2, a_3, ...)$, then $Sp_t = (0, a_1, a_2, ...)$.

Proof. The proof for Proposition 1 is presented in a constructive manner. That is, we build a Markov chain, X_t, with the desired properties. It is clear that the Markov chain should have the distribution p_0 at the initial state, at time zero. As mentioned earlier, this Markov chain can only have self-transitions, and transitions to the nearest neighbor to the right. Given these two conditions, we can derive the progression of the Markov chain over time. It is also worth reiterating that the Markov chain transition probabilities are governed by m, the number of messages already transmitted during the current transmission round, and t the current time slot. For the left edge of the Markov chain, when $m = 0$, we gather that

$$\begin{aligned} p_{t+1}(0) &= \Pr(X_{t+1} = 0 | X_t = 0)p_t(0) \\ &\triangleq \left(1 - \gamma_0^{(t)}\right)p_t(0) \end{aligned} \tag{3.3}$$

where γ_m^t is the conditional probability of a device transmitting its message during time slot t, given that it has already transmitted m messages. We now can write,

24

for any state m in the Markov chain,

$$p_{t+1}(m) = \Pr(X_{t+1} = m | X_t = m - 1)p_t(m - 1)$$

$$+ \Pr(X_{t+1} = m | X_t = m)p_t(m) \tag{3.4}$$

$$\triangleq \gamma_{m-1}^{(t)} p_t(m - 1) + \left(1 - \gamma_m^{(t)}\right)p_t(m).$$

These equations can be rearranged to arrive at the following:

$$\gamma_m^{(t)} p_t(m) = p_t(m) - p_{t+1}(m) + \gamma_{m-1}^{(t)} p_t(m - 1)$$

$$= \sum_{\ell=0}^{m} p_t(\ell) - \sum_{\ell=0}^{m} p_{t+1}(\ell) \tag{3.5}$$

where m takes value in $\mathbb{N}_0$.

Given that γ_m^t is a conditional probability, we can conclude that its value lies in $[0, 1]$. Thus, we can conclude two necessary and sufficient conditions. The first is that γ_m^t is non-negative, which is a result of the fact that γ_m^t is a conditional probability. Thus, we assert that the difference between the two sums in (3.5) needs to remain non-negative for all $m \in \mathbb{N}_0$. In terms of stochastic dominance, this translates to the condition that $p_t \preceq p_{t+1}$ must hold. The other requirement that can be inferred is that γ_m^t is no greater than one. This can be expressed as

$$\sum_{\ell=0}^{m} p_t(\ell) - \sum_{\ell=0}^{m} p_{t+1}(\ell) \leq p_t(m) \tag{3.6}$$

or, equivalently,

$$\sum_{\ell=0}^{m-1} p_t(\ell) \leq \sum_{\ell=0}^{m} p_{t+1}(\ell)$$

In terms of the standard right shift operator S, acting on one-sided infinite sequences of numbers, we can arrive at the expression $p_{t+1} \preceq Sp_t$. Thus, we see that the condi-

tions defined in (3.4) are necessary and sufficient to make possible the construction of time-dependent transition probabilities for a Markov chain such that the Markov chain expressed the desired distributions at every step. In addition, we can conclude that, for any Markov chain with satisfactory transition probabilities and only self-transitions and transitions to the nearest neighbor to the right, the sequence of probability distributions $p_0, p_1, \ldots$ that are achieved at every step must satisfy the stochastic dominance conditions defined in (3.4). $\qquad\square$

The partial order that is induced by stochastic dominance is essential in determining which distributions can be induced by the Markov chain structure that we have defined. Proposition 1 does not only provide necessary and sufficient conditions that must be satisfied by probability distributions in order to be represented by Markov chains of our defined form; it also provides a way to calculate the actual transition probabilities of the resulting Markov chain from the probability distributions.

3.2.1   Definition of Uncoordinated Markov Transmission Strategy

Let $p_0, p_1, \ldots$ be a series of probability distributions for which we desire to produce a Markov chain that admits these distributions at each step. In addition, let these probability distributions satisfy the stochastic dominance requirements defined in (3.4). Then, for integers $m, t \in \mathbb{N}_0$, we define parameters

$$\mu_t(m) = \gamma_m^{(t)} = \begin{cases} \frac{\sum_{\ell=0}^{m} p_t(\ell) - \sum_{\ell=0}^{m} p_{t+1}(\ell)}{p_t(m)} & p_t(m) > 0 \\ 0 & p_t(m) = 0. \end{cases} \qquad (3.7)$$

Let $\{X_t\}$ be a Markov chain with probability distribution p_0 in the initial state, and transition probabilities

$$\Pr(X_{t+1} = m | X_t = m) = 1 - \gamma_m^{(t)} \tag{3.8}$$

$$\Pr(X_{t+1} = m + 1 | X_t = m) = \gamma_m^{(t)} \tag{3.9}$$

where $m, t \in \mathbb{N}_0$. It is clear that $\Pr(X_{t+1} = n | X_t = m) = 0$ whenever $n \notin \{m, m+1\}$. This equates to a device only being able to send exactly zero or one messages during a time slot. As defined above, $\{X_t\}$, or equivalently $\boldsymbol{\mu}$, is a valid uncoordinated transmission policy.

Corollary 1. *The discrete random process $\{X_t\}$ specified above is a valid Markov chain and it possesses discrete distribution p_t at time t.*

Proof. The proof of Corollary 1 is largely a conclusion of Proposition 1. From the proof of Proposition 1, we have already shown that $p_t \preceq p_{t+1}$ necessitates that the transition probabilities defined in (3.7) are non-negative. In addition, we have already shown that $p_{t+1} \preceq S p_t$ guarantees that the transition probabilities will be less than one. From these two conditions, we have shown that the Markov chain X_t is a first-order Markov chain, and that it has initial distribution p_0. Also, through the equations (3.3) and (3.4), we have shown that the probability distribution of X_t at time slot t shapes the probability distribution p_t. Lastly, by its definition, the only transitions possible are self-transitions and transitions to nearest neighbors on the right. $\qquad\square$

3.3   Distribution Shaping Examples

In this section, we give examples of using the above described framework to shape Markov strategies for various distributions.

3.3.1 Soliton Distributions

For $t \in \mathbb{N}$, the probability distribution for the Soliton distribution is defined as

$$
p_{\text{sol}(t)}(m) = \begin{cases} \frac{1}{t} & m = 1 \\[2ex] \frac{1}{(m-1)m} & m = 2, \ldots t. \end{cases}
$$

We start by checking that the Soliton distributions satisfy the requirements of Proposition 1. For the first condition, we see that

$$
\sum_{\ell=0}^{m} p_t(\ell) - \sum_{\ell=0}^{m} p_{t+1}(\ell) = \frac{1}{t} - \frac{1}{t+1} = \frac{1}{t(t+1)}
$$

for $m = 0, \ldots, t$. And, we note that the difference vanishes for $m \geq t + 1$. This shows that the stochastic dominance requirement for the first condition stands, as $p_{\text{sol}(t)} \preceq p_{\text{sol}(t+1)}$. To verify the second condition in Proposition 1, we note that

$$
\sum_{\ell=0}^{m} p_{t+1}(\ell) - \sum_{\ell=0}^{m-1} p_t(\ell) = \frac{1}{t+1} \geq 0
$$

for $m = 1$; Next, we have

$$
\sum_{\ell=0}^{m} p_{t+1}(\ell) - \sum_{\ell=0}^{m-1} p_t(\ell) = \frac{1}{(m-1)m} - \frac{1}{t(t+1)} \geq 0
$$

for $m = 2, \ldots, t$. Again, for $m \geq t+1$ the distributions are equal to zero and, hence, $p_{\text{sol}(t+1)} \preceq S p_{\text{sol}(t)}$. Thus, the sequence of Soliton distributions satisfies both conditions of Proposition 1. Consequently, the distributions are an admissible sequence for an uncoordinated Markov strategy. For this example, consider the sequence

$$
p_0 = e_0, p_1 = p_{\text{sol}(1)}, \ldots, p_t = p_{\text{sol}(t)}, \ldots
$$

where e_0 is simply a unit vector with the mass at zero, indicating that we wish the chain to start at zero for all devices. We emphasize that $e_0 \preceq p_1$ and $p_1 = Se_0$, which satisfies the stochastic dominance conditions. From our previously defined Corollary 1 we can arrive, for these distributions, at the following transmission probabilities for the Markov strategy

$$
\gamma_m^{(t)} = \begin{cases} \frac{1}{t+1} & m = 1 \\ \frac{(m-1)m}{t(t+1)} & m = 2, \ldots, t \\ 0 & \text{otherwise.} \end{cases}
\tag{3.10}
$$

for $t \in \mathbb{N}$. The transition probabilities for the Markov chain $\{X_t\}$ for $t \in \mathbb{N}$ are given by

$$
\Pr(X_{t+1} = m | X_t = m) = 1 - \frac{(m-1)m}{t(t+1)}
\tag{3.11}
$$

$$
\Pr(X_{t+1} = m+1 | X_t = m) = \frac{(m-1)m}{t(t+1)}
\tag{3.12}
$$

for $m = 2, \ldots, t$; and

$$
\Pr(X_{t+1} = 2 | X_t = 1) = \frac{1}{t+1}
$$

$$
\Pr(X_{t+1} = 1 | X_t = 1) = \frac{t}{t+1}.
$$

At time zero, the Markov chain starts in state $X_0 = 0$ for all devices, and always transitions to $X_1 = 1$. This translates to each device transmitting at time slot 0. The transmission probability for a given device is subsequently given by

$$
\sum_{i=1}^{t} \gamma_i^{(t)} p_{\text{sol}(t)}(i) = \frac{1}{t+1} \quad \text{for } t \geq 2.
$$

29

It has been shown in [13] that the Soliton distribution is an optimal distribution for the number of times for each device to transmit when in the more traditional setting of slotted ALOHA with successive interference cancellation. However, there are some issues that make this distribution less than optimal in our problem formulation. The primary issue is that, although the Markov strategy achieves the desired Soliton distribution for the degree of the variable nodes, it does not achieve the desired distribution on the check nodes. More specifically, this Markov strategy tends to inform devices to transmit their messages far more often at the beginning of a transmission round than later on. This results in the ensuing bipartite graph not having the desired number of messages decoding after successive interference decoding. The following two distributions that we introduce are intended to address this issue.

3.3.2 Stateless Distributions

The stateless distribution is a distribution that determines the transition probabilities without considering the number of times that the device has already transmitted since the beginning of the round. That is, this distribution makes the decision of whether or not to transmit a device's message based only on the current time slot. The initial distribution for this stateless distribution is $p_0 = e_0$. The following distributions are defined recursively as follows,

$$p_{t+1} = \left(1 - \gamma^{(t)}\right)p_t + \gamma^{(t)} S p_t.$$

As before, S is simply the standard right shift operator. We consider two options for the emission probabilities,

$$\gamma^{(t)} = \frac{c}{t} \quad \text{and} \quad \gamma^{(t)} = 1 - \exp\left(\frac{c \log(\epsilon)}{t}\right) \tag{3.13}$$

where c is a tuning parameter. The first emission probability, $\gamma^{(t)} = \frac{c}{t}$ is designed to maximize the probability of getting a degree one check node near the end of the round. The benefit to this is clear, as degree one check nodes not only represent a decodable message, but also allow the decoding process to continue. The second emission probability, $\gamma^{(t)} = 1 - \exp\left(\frac{c \log(\epsilon)}{t}\right)$ results in reducing the probability of an empty slot at the end of the round being reduced to ϵ. This is beneficial because empty slots provide no benefit to the decoding process.

3.3.3 Skewed Distributions

The skewed distributions are designed to balance the effects of the stateless distributions described above. Rather than simply relying on the information of the current time slot to decide transition probabilities, these distributions primarily rely on the number of times that a device has transmitted. More specifically, these distributions make it more likely for devices to transmit their message if they have already transmitted their message many times. The intended effect is that devices that have transmitted often will have a higher chance of being decoded, and will serve the decoding process well by removing their message from many collisions. As with the stateless distributions, the skewed distributions start with an initial distribution $p_0 = e_0$. With an emission target $\overline{\gamma}^{(t)}$, the state transition probabilities at time t are

defined as

$$
\gamma_m^{(t)} = \begin{cases} 0, & \sum_{i=0}^{m} p_t(i) < 1 - \overline{\gamma}^{(t)} \\ 1, & \sum_{i=m}^{t} p_t(i) \leq \overline{\gamma}^{(t)} \\ \frac{\overline{\gamma}^{(t)} - \sum_{i=m+1}^{t} p_t(i)}{p_t(m)} & \text{otherwise.} \end{cases}
$$

A slightly different version of the skewed distributions puts a limit on the number of packets that a device can send by any time slot t.

Both of these distributions are based on the same intuition: balancing the likelihood of empty slots with the probability of getting single check nodes near the end of a transmission round. Both of these are problems with the Markov strategy that is shaped to emit the Soliton distribution. In our results section, we show that a mixture of the skewed and stateless distributions can yield very promising results in terms of the throughput.

4. NUMERICAL RESULTS*

In this section, we discuss our numerical results, as well as the methods by which we came to these results. We also analyze the numerical results and draw conclusions where appropriate.

4.1 Simulations

For our numerical results, we have written a framework for simulating a transmission round for devices employing our Markov strategies, as well as the resulting decoding process. Our framework is written in Python 2. The simulations are run as follows. For a given K devices, N time slots, and a Markov strategy X_t, the simulation generates an equivalent bipartite graph for a transmission round. The bipartite graph is then run through a successive interference decoding process, as explained in Algorithm 1; and the number of decoded messages is noted. From this, we can infer the throughput of the simulated round. It is worth noting that our simulation makes no attempt to actually send messages or perform actual message decoding in signal space. Rather, it generates a bipartite graph, where an edge between a variable node v and check node c means that the device denoted by v transmits a copy of its message during the time slot represented by c. The following numerical results are taken from running a large number of realizations for each case. As discussed earlier, the throughput denoted in these results is defined as the average number of messages decoded per time slot.

4.2 Results

In Figure 4.1, we see the throughput results for the Soliton distributions, as shaped by a Markov strategy. The process for shaping a Markov strategy for the Soliton distributions, and the actual transition probabilities are detailed in Chapter 3. In the simulations used to arrive at these results, the number of devices is $k = 1000$. The value for the number of time slots, denoted n, varies from 0 to 2000. For each number of time slots, a number of samples are run and the average throughput is computed. As discussed earlier, the Soliton distributions as shaped by the Markov framework are less than ideal. It has been shown that the Soliton distribution is ideal in the more traditional slotted ALOHA with successive interference cancellation decoding. Although we see the correct marginal distribution shaped for the degrees of the variable nodes in the resulting bipartite graph, this does not induce the desired distribution on the check node side. The earlier time slots in the transmission round are heavy with collisions, and the later time slots are often empty. As a result, the throughput performance of the Soliton distributions fails to meet our expectations. To be exact, we see a peak throughput performance of around 52 percent. The curves shown in Figures 4.2, 4.3, 4.4, and 4.5 are the results for a distribution sequence that is a mixture of the skewed and stateless distributions. The simulations run for these figures are identical to the simulations for Figure 4.1, with the Markov strategy replaced with a Markov strategy shaping the skewed and stateless mixture distributions. The plots are for a fixed number of devices of 250, 500, 1000, and 2000. As discussed earlier, these distributions are designed to alleviate the issue of having an unbalanced distribution on the check nodes of the resulting bipartite graph. The value of c in seen for the different curves in the figure is a parameter to the distribution, determining some features of the mixture of the distributions.
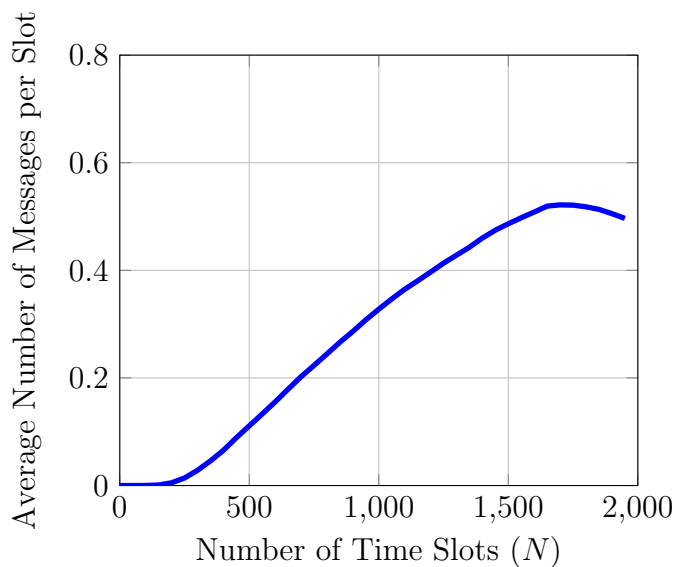
Figure 4.1: This graph showcases overall throughput performance, in terms of decoded messages per slot, for $K = 1000$ active devices using successive interference cancellation at the access point.

For these distributions, we see a much higher peak throughput than for the Soliton distributions, which is to be expected. In particular, we see a maximum throughput of slightly greater than 70 percent, with $N = 2500$ time slots and $c = 1.2$, for the figure for the number of devices being fixed at 2000. We note that we see slight throughput performance increases as the number of devices is increased. This result is promising because the throughput significantly exceeds the performance of traditional ALOHA. These results show that the uncoordinated Markov strategy can perform to desirable levels. Nevertheless, the throughput shown by our simulations with the mixed distributions does not meet the throughput shown possible in more traditional uses of slotted ALOHA with successive interference cancellation. This is expected as our problem formulation is more restrictive and is less coordinated.
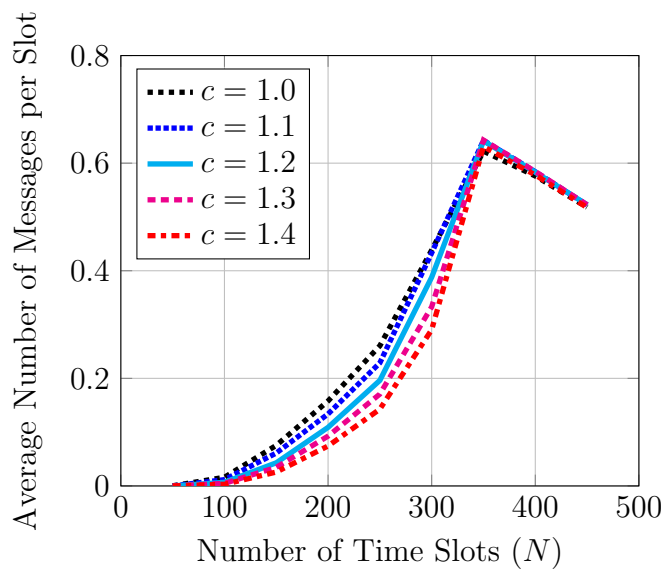
Figure 4.2: This graph presents overall throughput performance for $K = 250$ devices.
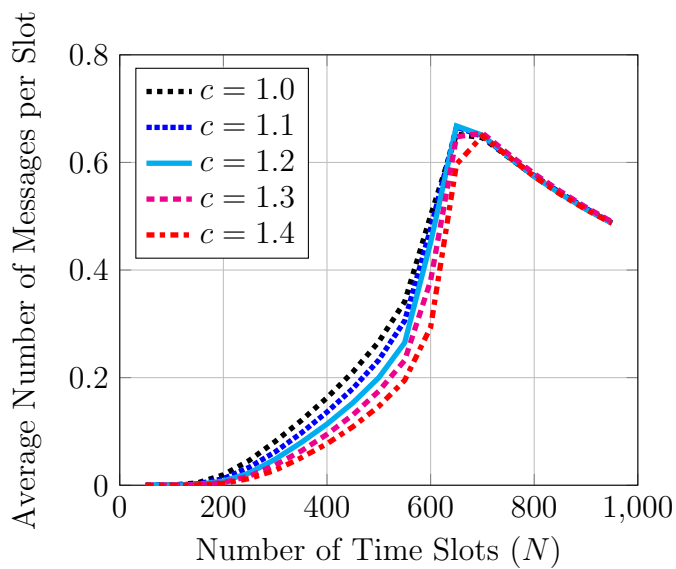

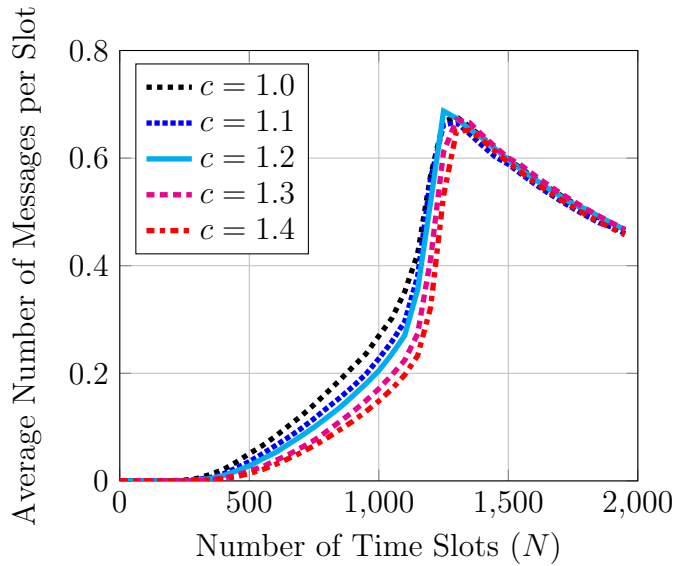Figure 4.3: This graph presents overall throughput performance for $K = 500$ devices.

Figure 4.4: This graph presents overall throughput performance for $K = 1000$ devices. Reprinted with permission from "On the design of universal schemes for massive uncoordinated multiple access" by Austin Taghavi, Avinash Vem, Krishna Narayanan, and Jean-Francois Chamberland. 2016. ISIT July 2016: 345-349 by IEEE.
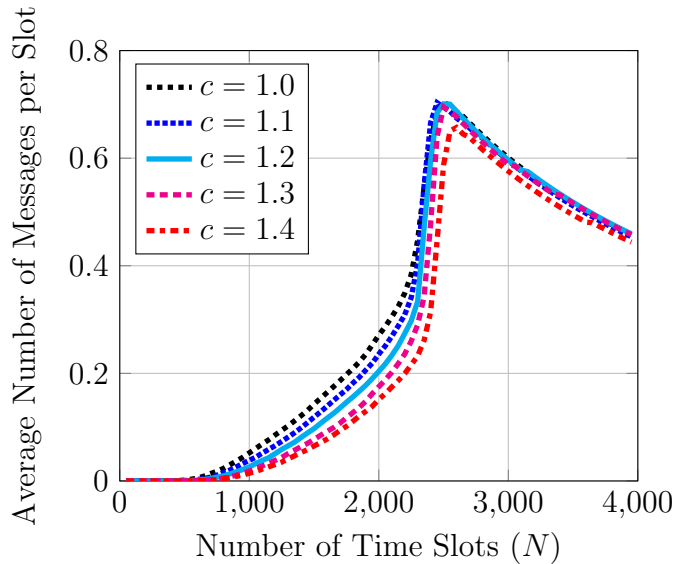

Figure 4.5: This graph presents overall throughput performance for $K = 2000$ devices.

5.   MULTIPLE ACCESS POINTS

5.1   System Model

In realistic implementations of systems requiring a random access scheme, there is often more than just one access point. If multiple access points are used to cover an area and have overlapping footprints, then it is feasible that some of the devices are overlapping devices in the sense that they will be transmitting their message to multiple access points at one time. To simplify this line of thinking, let us consider the case of two access points with overlapping devices. Figures 5.1, 5.2, and 5.3 show possible diagrams for such a model. In each diagram, each point represents an active device in the network, while the triangles are access points and their footprints are shaded in blue. The green points are overlapping devices, and the black points are not. We will refer to the left most access point as access point A, and the right most as access point B. Figure 5.1 shows a case where the footprint of the access points is not large enough to cover the devices. While there is no overlap, there are also points who cannot transmit their message to either access point. Figure 5.2 shows a scenario where the footprint of the access points is too large. In this case, the devices would largely be transmitting to both access points, which would be very inefficient. Figure 5.3 shows a feasible diagram for a two access point slotted ALOHA based communication scheme. In this diagram, there are four overlapping devices, and four non-overlapping devices. All devices are covered by access point A, access point B, or both. As explained in [9], the results of a transmission round for random access schemes can be represented with a bipartite graph. For the multiple access point scenario depicted in Figures 5.1, 5.2, and 5.3, the analogous representation would be two connected bipartite graphs. The shared nodes between the two graphs
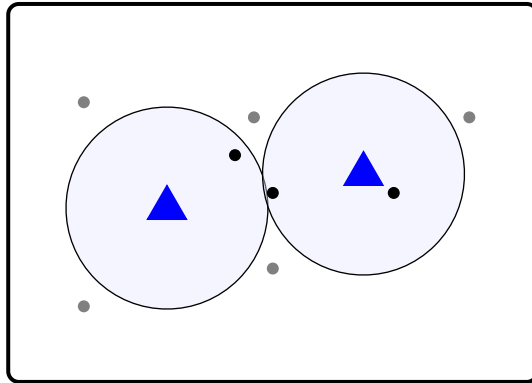
38

Figure 5.1: A diagram showing a possible multi access point scenario in which not enough devices are reached.
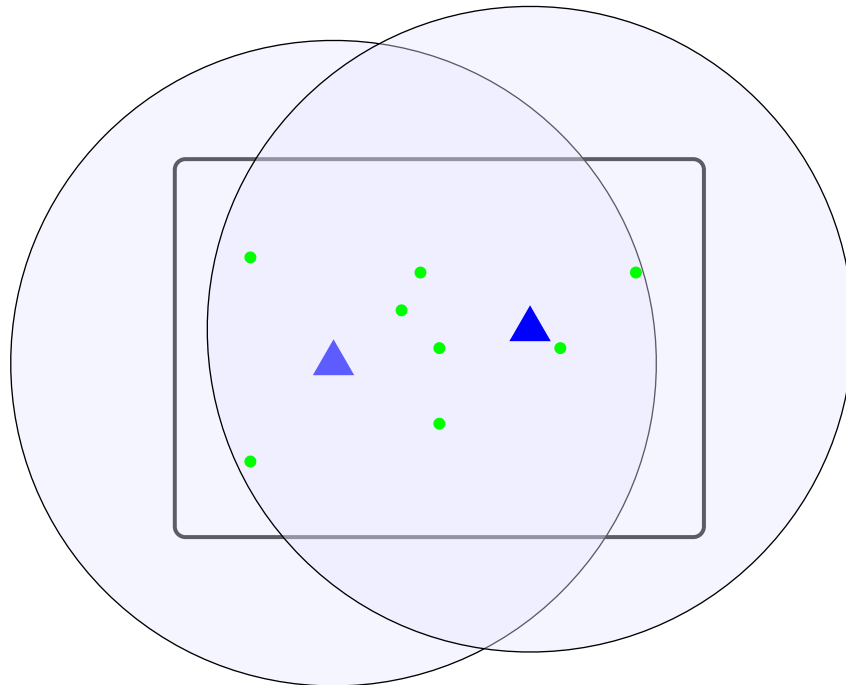

Figure 5.2: A diagram showing a possible multi access point scenario in which the footprints of the access points are too large.
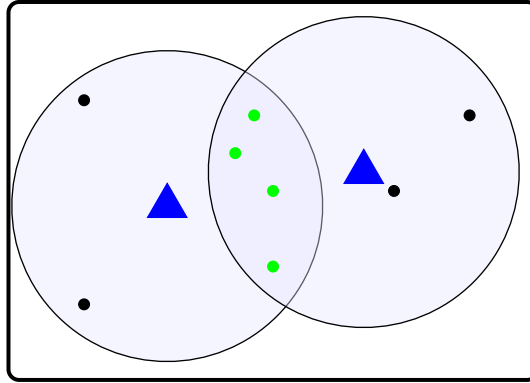
Figure 5.3: A diagram showing a feasible multi access point scenario with two access points.

would be the overlapping devices. The decoding process for such a graph is similar to the standard successive interference cancellation. However, after each iteration of decoding, the decoded nodes can be shared with the other side of the graph to attempt to continue the decoding process. This is simply saying that if access point A can decode the message device x, and device x is overlapping with access point B, then the message for device x can be used in the interference cancellation process for access point B. In other words, the access points have an alternate means to shared decoded messages. We believe that this is a reasonable amount of information to share between the access points. Only the decoded packets are intended to be shared between access points, rather than the entire signal that is received. While sharing the entire signal between access points is likely too much information to be realistically shared, the decoded packet information is orders of magnitude smaller. It is worth noting that we are considering this problem formulation as a case of the more classical slotted ALOHA with successive interference cancellation, in that we may know the number of users in the system.

5.1.1   Connected Bipartite Graphs

In this section, we show some possible resulting graphs for the two access point problem formulation. Note that these graphs are not simply bipartite graphs, but rather *connected* bipartite graphs. We seek to define the connected bipartite graph in the simplest terms possible. First, a connected bipartite graph must contain two sets of analogous check nodes. That is, the two sets of check nodes C_A and C_B have the same cardinality, and for each element c_A in C_A there exists a paired element c_B in C_B. In the application of this bipartite graph to our problem formulation, two check nodes are paired if they represent the same time slot on the two access points. A connected bipartite graph must also contain three sets of variable nodes V_A and V_B, and V_O. The sets $\{V_A, C_A\}$ and $\{V_B, C_B\}$ form valid bipartite graphs. The elements in V_O may have edges to C_A and C_B. However, for each edge between v_O in V_O and c_A in C_A, there must also be an edge connecting v_O and c_B, where c_B is the paired node of c_A. In our use of the graph, the elements in C_A and C_B represent the time slots for access point A and access point B, respectively. The elements in V_A and V_B are the devices that transmit to, and only to, access points A and B, respectively. The elements in V_O represent the devices that simultaneously transmit to both access point A and access point B, i.e., the set of overlapping devices. A connected bipartite graph representing a possible transmission round for the scenario depicted in Figure 5.3 is shown in Figure 5.4. We consider the leftmost access point in Figure 5.3 to be access point A, and the rightmost access point to be access point B. For this transmission round, only 4 time slots per access point were used in the round. In this graph, the time slots labeled "slot A 1" through "slot A 4" represent time slots for access point A, and the time slots labeled "slot B 1" through "slot B 4" represent time slots for access point B. Devices 1 and 2 are the leftmost devices
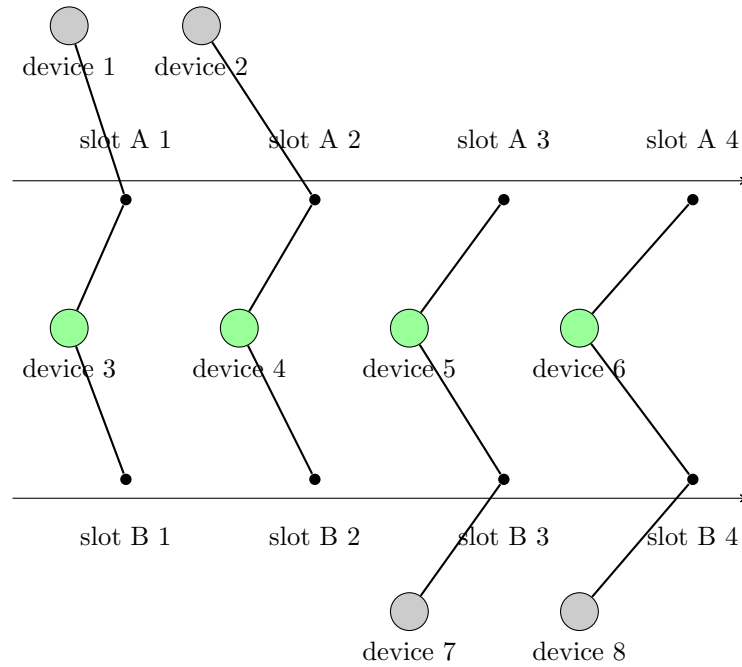
Figure 5.4: A diagram showing a connected bipartite graph representing a possible transmission round for Figure 5.3.
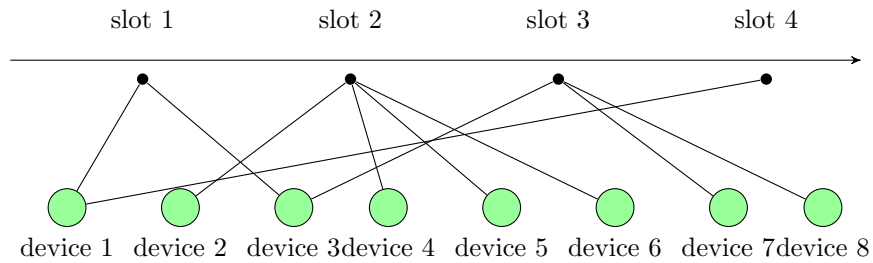

Figure 5.5: A diagram showing a connected bipartite graph representing a possible transmission round for Figure 5.2.

in Figure 5.3 , which are only visible to access point A. Similarly, devices 7 and 8 are the rightmost devices that are only visible to access point B. Devices 3, 4, 5, and 6 are the devices in Figure 5.3 which are visible to both access point A and access point B. In this example bipartite graph, devices 1 and 2 represent V_A, devices 7 and 8 represent V_B, slots "A 1" through "A 8" represent C_A, slots "B 1" through "B 8" represent C_B, and devices 3, 4, 5, and 6 represent V_O. The paired check nodes are $\{(A1, B1), (A2, B2), \ldots, (A8, B8)\}$

Figure 5.5 shows a possible connected bipartite graph for the scenario depicted in Figure 5.2 . It is clear that the decoding process on this graph will largely be unsuccessful. In order for such a graph to be decodable, more time slots would be needed. In this case, the connected bipartite graph is simply two copies of the same bipartite graph, which is clearly undesirable.

5.1.2   Preliminary Observations

For this section, we will consider a problem formulation with $2k$ total active devices, and two access points. We will denote the number of users transmitting to only access point A as u_B, those only transmitting to access point B as u_B, and those transmitting to both as u_O. We note that it seems ideal for each access point to reach exactly k active devices with its footprint, with no devices overlapping. In a traditional CRDSA type setting, this would simply reduce to two separate schemes with k users each, and we could decode all of the messages with a little more than k time slots. However, we want to consider the case where there exist a set of overlapping users, as it is realistic that this scenario will arise. We want to determine cases where we can achieve the same efficiency as the above described case, but with a set of overlapping users.

First, we consider one extreme case: $2k$ overlapping users. In this case, all users

transmit to both access points. Thus, we have an equivalent scenario to a single access point with $2k$ users. We know that this scheme will need $2k$ or more time slots to successfully decode all of the messages, which is not the desired efficiency. We can actually gather that any scheme with some $u_O > k$ overlapping users will take at least u_O time slots to decode. We also note that any scheme with some $u_A > k$ or $u_B > k$ will also take greater than k time slots to decode all messages.

From this, we can conclude that our ideal layout must have the following properties: $u_A \leq k$, $u_B \leq k$, and $u_O \leq k$. We also note that, necessarily, $u_A + u_B + u_O = 2k$. It is worth noting that if any of $\{u_A, u_B, u_O\}$ are equal to k, then the other two will be necessarily less than or equal to k. Furthermore, if u_A or u_B is equal to k, then we can decode all of the messages in k time slots with traditional distributions. Consider the case where $u_A = k$. Then, $u_B + u_O = k$. So, the bipartite graph produced by access point B contains k variable nodes and k check nodes. This graph has been shown to be largely decodable with the devices using a Soliton degree distribution. The bipartite graph for access point A has $u_A + u_O$ variable nodes, and k check nodes. After access point B has decoded all of its messages, the information can be passed to access point A for decoding, leaving access point A with only $u_A = k$ variable nodes left to decode. Again, if these devices transmit according to the Soliton distribution, this graph will be decodable as well.

It is less clear what will happen in the case of $u_A + u_O > k$, $u_B + u_O > k$, $u_O \leq k$, $u_A \leq k$, and $u_B \leq k$. We show in the following section that it is possible to have a throughput of 1.0 even if $u_O = k$. However, the best distributions for such a scenario are not obvious. While these observations do not answer all of the relevant questions about this problem formulation, they do provide some bounds and some clarity on certain subsets of the problem, and identifies some subsets that are interesting to work with.

5.2 Decoding

The decoding algorithm for the two access point problem formulation an be described as follows.

Algorithm 2 Successive Interference cancellation For Two Access Points

---
  procedure PEELING DECODER FOR TWO ACCESS POINTS, A AND B
      Degree1NodesA ← Nodes of degree one from A
      Degree1NodesB ← Nodes of degree one from B
      while Degree1NodesA.size() + Degree1NodesB.size() > 0 **do**
          Run peeling decoder round for A, using Degree1NodesA ∪ Degree1NodesB
          Run peeling decoder round for B, using Degree1NodesA ∪ Degree1NodesB
          Degree1NodesA ← Nodes of degree one from A
          Degree1NodesB ← Nodes of degree one from B

5.2.1 Peeling Decoder Example

We can see that the connected bipartite graph shown in Figure 5.4 can be fully decoded. First, we note that slots "A 3", "A 4", "B 1", and "B 2" are all degree one. From these time slots, we can decode the messages for devices 3, 4, 5 and 6. Once the edges for devices 3, 4, 5, and 6 are removed from the graph, we see that the remaining time slots are all degree one. From this remaining graph, devices 1, 2, 7, and 8 are all decodable. While this is simply an ideal example, it does demonstrate the usefulness of the shared decoding process. Without sharing information between the two access points during the decoding process, devices 1, 2, 7, and 8 would not be decodable. The benefit of shared decoding in this example is a throughput increase from 0.5 to 1.0.
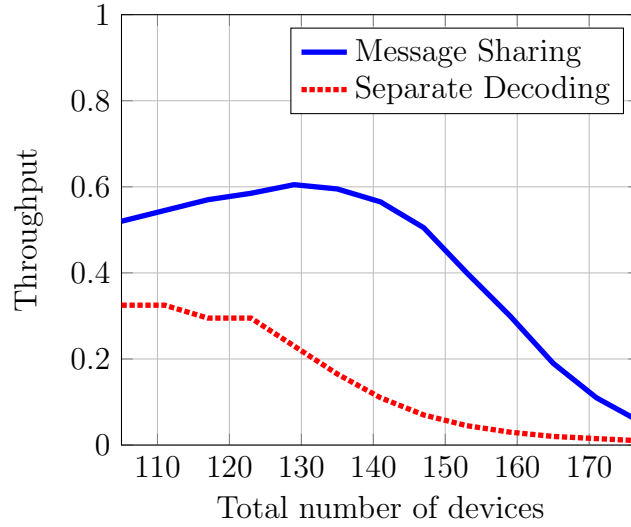
Figure 5.6: Performance for 2 receivers with N=100 time slots.

5.3   Numerical Results

Figure 5.6 shows the results of a two access point simulation, with an even $1/3$ of the active devices in each of u_A, u_B, and u_O. The graph shows the throughput in relation to the total number of devices in the system. The purpose of this result is to demonstrate the efficiency that the sharing decoding process adds to the scheme. It is clear that sharing information about decoded overlapping devices can lead to a more efficient decoding process. This gives us promise that even more efficient schemes can be designed specifically for the multiple access point models. It is our intuition that optimal schemes for such a problem formulation exist and can be derived. Our research has only scratched the surface of this problem by defining the problem formulation and numerically exploring the performance of the shared decoding process.

6. CONCLUSIONS

The ALOHA protocol has been a staple of wireless communication since its introduction in 1971, in particular in the fields of satellite communication and RFID identification. As other protocols and technologies have evolved, researchers have improved and modified ALOHA and slotted ALOHA to meet the performance standards for modern communication. In particular, the relation between slotted ALOHA communication and erasure coding theory, and the representation of a slotted ALOHA transmission round as a bipartite graph, has opened an entire branch of research to be applied to random access schemes.

One particular condition of slotted ALOHA problem formulations has largely been present in previous research: the number of active devices in the model is known or estimated. However, this is not always ideal, in particular with the changing landscape of modern communication. As IoT networks become more prevalent, the number of active devices in such systems may be hard or impossible to estimate, or simply may change frequently. In our research, we have considered the problem formulation of slotted ALOHA with successive interference cancellation, as in CRDSA [2], but with the added constraints that the number of users is not known or estimated, and the number of time slots in the round is not decided upon a priori. This problem formulation is not only novel, but very applicable to the current mobile communication landscape.

The first contribution of this research has been to provide a clear and complete problem formulation for slotted ALOHA with successive interference cancellation decoding, that is agnostic to the number of users in the system and does not decide on the number of time slots a priori. We defined the conditions and constraints of the

model, and explained why the model is novel. In addition, we examined the reasons that more traditional approaches can not be applied to this model. In particular, we showed the need for a transmission policy that can form the desired distributions without needing to know the total number of users in the system or the total number of time slots in the transmission round.

The second contribution of this research has been to provide a framework to creating transmission policies that can shape the desired distributions for the problem formulation that we described. We found a first-order Markov strategy over the set $\mathbb{N}_0$, with only self-transitions and unit transitions to the right, to be a suitable way to represent such a transmission policy. We derived necessary and sufficient conditions for a set of distributions to be suitable to be shaped by such a first-order Markov strategy, and gave a clear definition of an uncoordinated Markov strategy. Finally, we showed 3 examples of uncoordinated Markov strategies and their transition probabilities, shaped to represent specific distributions. Two of these distribution, the skewed and stateless distributions, were designed specifically for this problem formulation.

Our numerical results showed that the throughput of a scheme that employs an uncoordinated Markov strategy shaping a Soliton distribution is less than ideal. This is because the distribution on the check nodes is concentrated on the time slots towards the beginning of the transmission round, and leaves more empty slots at the end of the transmission round. However, a Markov strategy that shapes mixture of the skewed and stateless distributions gives very promising results. While the throughput is not as high as for traditional CRDSA, the results are still satisfactory. Our problem formulation should be expected to have a lower throughput, as the structure of the communication is less coordinated. These results show that a random access scheme can achieve a reasonable throughput without needing to know or estimate the number of active users in the system. This is encouraging, as it provides

systems that cannot know the number or users or accurately estimate the number of users a way to use random access schemes with high efficiency.

There are some questions that are left unanswered by our research that would certainly be interesting to explore. First, we have not found a throughput limit for random access schemes using uncoordinated Markov strategies, either numerically or analytically. While we have defined distributions and shown numerically that the throughput performance is promising, we have yet to define an optimal distribution. It would be very beneficial to derive the optimal uncoordinated Markov strategy and show the expected throughput in the limit for a large number of devices, as was done for slotted ALOHA with successive interference cancellation in [13]. This research has shown that it is possible to achieve high throughput in the novel problem formulation that we defined, but it leaves open the question of what the optimal strategy looks like.

6.1   Multiple Access Points

Our research also considers a different, largely unexplored, problem formulation for random access schemes. In this problem formulation, there exist two access points to which the active devices wish to transmit their messages. Active devices are in the footprint of either one or both of the access points. If an active devices is in the footprint of both of the access points, it transmits its message to both access points during time slots in which it chooses to transmit. We give a problem formulation for this novel scenario, and explain why this scenario is reasonable and could be applicable to some implementations. In addition, we give a definition for the structure of the resulting graph representation of a transmission round for this problem formulation. The structure can be thought of as two connected bipartite graph. We also describe the successive interference decoding process for the connected bipartite graph, which

relies on the access points sharing information during the decoding process. Our results show that the decoding process we describe can be very efficient, and gives an enormous increase in performance over a naive method of trying to decode the graphs separately.

This problem formulation has implications for antenna design, as well as mobile communication. If the structure of this graph and optimal distributions for the degrees of nodes are well studied, then this research can be used to inform optimal layouts and footprints for antennas used for such communication schemes.

There are many unexplored avenues for this branch of our research. A productive route would be to extend what we have done for two access points to n access points. This includes a definition of the problem formulation, as well as the graph structure of a transmission round, and the decoding process for such a graph. The graph structure and decoding process will increase in complexity as the number of access points grows. This would certainly be useful, as realistic implementations could conceivably consist of any number of access points with overlapping footprints. It is also still unclear which distributions are optimal for this problem formulation, and whether a different transmission policy for overlapping devices is desirable or beneficial. It is clear that decoding messages that are present on both access points is particularly helpful in the decoding process, as it provides benefits to the decoding process for both access points. However, the ideal degree distributions for the connected bipartite graph is left to be determined.

REFERENCES

[1] Norman Abramson. The ALOHA system: Another alternative for computer communications. In *Proc. Computer Conference*, pages 281–285. ACM, 1970.

[2] Enrico Casini, Ricardo De Gaudenzi, and Oscar Del Rio Herrero. Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access scheme for satellite access packet networks. *IEEE Transactions on Wireless Communications*, 6(4):1408–1419, April 2007.

[3] Gagan Choudhury and Stephen Rappaport. Diversity ALOHA–a random access scheme for satellite communications. *IEEE Transactions on Communications*, 31(3):450–457, 1983.

[4] M. Scott Corson and Anthony Ephremides. An analysis of multi-receiver, non-adaptive, slotted ALOHA with capture for wireless communications in factories. In *INFOCOM'93. Proceedings. Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking: Foundation for the Future, IEEE*, pages 421–428. IEEE, 1993.

[5] Oscar del Rio Herrero. A high-performance mac protocol for consumer broadband satellite systems. *IET Conference Proceedings*, pages 512–512(1), January 2009.

[6] Mikhail Ivanov, Fredrik Brännström, Alexandre Graell i Amat, and Petar Popovski. Error floor analysis of coded slotted ALOHA over packet erasure channels. *IEEE Communications Letters*, 19(3):419–422, 2015.

[7] Dušan Jakovetić, Dragana Bajović, Dejan Vukobratović, and Vladimir Crnojević. Cooperative slotted ALOHA for multi-base station systems. *IEEE Trans-*

actions on Communications, 63(4):1443–1456, 2015.

[8] Su-Ryun Lee, Sung-Don Joo, and Chae-Woo Lee. An enhanced dynamic framed slotted ALOHA algorithm for rfid tag identification. In *Mobile and Ubiquitous Systems: Networking and Services, 2005. MobiQuitous 2005. The Second Annual International Conference on*, pages 166–172. IEEE, 2005.

[9] Gianluigi Liva. Graph-based analysis and optimization of contention resolution diversity slotted ALOHA. *IEEE Trans. Commun.*, 59(2):477–487, 2011.

[10] Michael Luby. Lt codes. In *Symp. Found. Comp. Sci*, pages 271–280. IEEE, 2002.

[11] David JC MacKay. Fountain codes. *IEE Proceedings-Communications*, 152(6):1062–1068, 2005.

[12] Andrea Munari, Michael Heindlmaier, Gianluigi Liva, and Matteo Berioli. The throughput of slotted ALOHA with diversity. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 698–706. IEEE, 2013.

[13] Krishna R. Narayanan and Henry D. Pfister. Iterative collision resolution for slotted ALOHA: An optimal uncoordinated transmission policy. In *Intern. Symp. Turbo Codes & Iter. Inf. Proc. (ISTC)*, pages 136–139. IEEE, 2012.

[14] Enrico Paolini, Cedomir Stefanovic, Gianluigi Liva, and Petar Popovski. Coded random access: How coding theory helps to build random access protocols. *arXiv preprint arXiv:1405.4127*, 2014.

[15] Michele Zorzi. Mobile radio slotted ALOHA with capture, diversity and retransmission control in the presence of shadowing. *Wireless Networks*, 4(5):379–388, 1998.