An Operations Research Approach to Aviation Security

by

Susan Elizabeth Martonosi

B.S., Cornell University, 1999

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2005

Author .....
Sloan School of Management
11 August 2005

Certified by .....
Arnold I. Barnett
George Eastman Professor of Management Science
Thesis Supervisor

Accepted by .....
James B. Orlin
Edward Pennell Brooks Professor of Operations Research
Co-Director, MIT Operations Research Center

An Operations Research Approach to Aviation Security

by

Susan Elizabeth Martonosi

Submitted to the Sloan School of Management
on 11 August 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Since the terrorist attacks of September 11, 2001, aviation security policy has remained a focus of national attention. We develop mathematical models to address some prominent problems in aviation security.

We explore first whether securing aviation deserves priority over other potential targets. We compare the historical risk of aviation terrorism to that posed by other forms of terrorism and conclude that the focus on aviation might be warranted.

Secondly, we address the usefulness of passenger pre-screening systems to select potentially high-risk passengers for additional scrutiny. We model the probability that a terrorist boards an aircraft with weapons, incorporating deterrence effects and potential loopholes. We find that despite the emphasis on the pre-screening system, of greater importance is the effectiveness of the underlying screening process. Moreover, the existence of certain loopholes could occasionally *decrease* the overall chance of a successful terrorist attack.

Next, we discuss whether proposed explosives detection policies for cargo, airmail and checked luggage carried on passenger aircraft are cost-effective. We define a threshold time such that if an attempted attack is likely to occur before this time, it is cost-effective to implement the policy, otherwise not. We find that although these three policies protect against similar types of attacks, their cost-effectiveness varies considerably.

Lastly, we explore whether dynamically assigning security screeners at various airport security checkpoints can yield major gains in efficiency. We use approximate dynamic programming methods to determine when security screeners should be switched between checkpoints in an airport to accommodate stochastic queue imbalances. We compare the performance of such dynamic allocations to that of pre-scheduled allocations. We find that unless the stochasticity in the system is significant, dynamically reallocating servers might reduce only marginally the average waiting time.

Without knowing certain parameter values or understanding terrorist behavior, it can be difficult to draw concrete conclusions about aviation security policies. Nevertheless, these mathematical models can guide policy-makers in adopting security measures, by helping to identify parameters most crucial to the effectiveness of aviation security policies, and helping to analyze how varying key parameters or assumptions can affect strategic planning.

Thesis Supervisor: Arnold I. Barnett
Title: George Eastman Professor of Management Science

Acknowledgments

This thesis is the product not just of my own work but also that of many who have helped me along the way. First, I am indebted to my thesis advisor, Arnold Barnett, for the wisdom and patience he has exhibited with me. His constant reminders to think simply when modeling a problem helped me out of many ruts, and I hope to remember this wise rule throughout my career. A talented writer and orator, he offered many useful suggestions and criticisms on the writing and presentation of this thesis, and this instruction will serve me well into the future. I also admire his commitment to problems of social importance. It was indeed this commitment that led me to work with him, and his emphasis on selecting methodology to suit the problem rather than the reverse has inspired me to continue in this domain.

Professors Cynthia Barnhart and Amedeo Odoni, who served on my thesis committee, have offered support and guidance since well before the official formation of the committee. In particular, their suggestions helped to clarify the focus of the problem posed in Chapter 5 and provided fruitful directions of exploration. Their kindness and encouragement throughout my time here will remain a lasting memory for me.

I am grateful for the generous support of the Alfred P. Sloan Foundation, via the MIT Global Airline Industry Program, who funded this work. Anthony Ventresca of Massport was also of great assistance in sharing data and expertise that were invaluable to the research in Chapter 5.

The Operations Research Center has become a second home to me, thanks to the efforts of the co-directors (Jim Orlin and John Tsitsiklis) and the administrative staff (Paulette Mosley, Laura Rose, Veronica Mignot and Andrew Carvalho) to make it a welcoming environment in which to work. I thank my fellow students for conversations, camaraderie, friendship, and, most of all, karaoke. A special thanks to those in my incoming class with whom I completed my coursework and studied for the qualifying exams. I learned so much by working with you.

Outside of the ORC, I wish to thank Rambax MIT for being my African oasis here in Boston, and my students in Koundara, Guinea for inspiring me with their determination.

Lastly, none of this would have been possible without the encouragement, unshakable confidence, and love of my family and God. It is the latter of these gifts for which I am most grateful.

Contents

1 Introduction **15**

2 The Relative Risk of Aviation Terrorism **21**
2.1 A relative risk factor 22
2.2 Terrorism against Americans, 1968-2001 23
2.2.1 Definition of terrorism 23
2.2.2 Classification of terrorist attacks 23
2.3 Fraction of time spent in aviation 24
2.3.1 Estimation based on passenger enplanements 25
2.3.2 Estimation based on time-use studies 26
2.4 Death risk per hour 26
2.5 Using the past to predict the future 27

3 How Effective Might Passenger Profiling Be? **29**
3.1 System and parameter description 31
3.2 Opposing viewpoints 34
3.2.1 The "right answer" to security? 34
3.2.2 A carnival game? 35
3.3 A general model 36
3.3.1 How terrorists might probe the system 36
3.3.2 Modeling the probability of a successful attack 37
3.3.3 Updating conditional probabilities of selection 39
3.4 Analysis techniques 42
3.5 Interpreting the model 44
3.5.1 Effects of screening effectiveness parameters on attack success 44
3.5.2 Deterrence effects 49
3.5.3 Role of random screening 54
3.6 Conclusions and policy implications 55

4 Securing Cargo Holds of Passenger Aircraft: A Cost-Benefit Perspective **57**
4.1 Three security measures 58
4.1.1 Positive Passenger-Bag Match 58
4.1.2 Moratorium on larger Postal Service airmail packages 59

4.1.3 A hypothetical policy on cargo . . . 60
4.2 A cost-benefit model . . . 60
4.2.1 Parameters and constants . . . 61
4.2.2 Model based on cost per flight . . . 62
4.2.3 A renewal model . . . 62
4.2.4 Interpretation . . . 64
4.3 Parameter Estimation . . . 66
4.3.1 The cost of attack . . . 66
4.3.2 Policy costs . . . 67
4.3.3 Backup security effectiveness . . . 69
4.3.4 Backup security costs . . . 71
4.4 Comparing the cost-effectiveness of cargo-hold measures . . . 71
4.5 Diversion of threats . . . 76
4.6 Conclusion . . . 78

5 Dynamic Security Screener Allocation **79**
5.1 Problem description . . . 80
5.2 Literature review . . . 81
5.3 Problem formulations . . . 85
5.3.1 Deterministic fluid model . . . 89
5.3.2 Deterministic disruptions to passenger entry pattern . . . 95
5.3.3 Stochastic disruptions to passenger entry pattern . . . 97
5.3.4 Randomized entry rates . . . 98
5.3.5 Stochastic service times . . . 99
5.4 Approximate dynamic programming solution techniques . . . 99
5.4.1 Aggregated state space with limited lookahead . . . 100
5.4.2 Approximating expected values . . . 101
5.5 Data . . . 103
5.6 Results . . . 106
5.6.1 Deterministic entries and services . . . 106
5.6.2 Deterministically disrupted entries . . . 110
5.6.3 Stochastically disrupted entries . . . 112
5.6.4 Randomized entry rates . . . 115
5.6.5 Stochastic service times . . . 116
5.6.6 General observations . . . 117
5.6.7 Sensitivity analysis . . . 123
5.7 Model limitations . . . 127
5.8 Are dynamic allocations of servers beneficial? . . . 129

6 Conclusions **131**

A Fatal Terrorist Attacks Against American Civilians, 1/1/1968 - 9/10/2001 **135**

B Logistic Regression of Dynamic Server Allocations 139

List of Figures

3-1 Current security checkpoint procedure. 32
3-2 The path through airport passenger screening for an individual terrorist carrying weapons 34
3-3 A decision tree describing the probability of a successful attack. 37
3-4 The average probability of a successful attack attempt, $P(AS)$, by (p_1, p_2) range 45
3-5 The minimum profiling effectiveness (C) required to experience an average $P(AS)$ less than or equal to 20%, 10% or 5%, by (p_1, p_2) grouping. 48
3-6 The probability of a successful terrorist attempt by deterrence threshold τ . 49
3-7 Probing the system can sometimes discourage rather than reassure the terrorist. 50
3-8 The probability, $P(AS)$ of a successful terrorist attempt, by random screening percentage, r. 55

4-1 PPBM: Time threshold for which the expected value of the policy is positive, versus attempt likelihood. 73
4-2 Airmail: Time threshold for which the expected value of the policy is positive, versus attempt likelihood. 73
4-3 Cargo: Time threshold for which the expected value of the policy is positive, versus attempt likelihood. 74
4-4 PPBM: 50% threshold time to first attempt by attack cost and backup security effectiveness. 75
4-5 Airmail: 50% threshold time to first attempt by attack cost and backup security effectiveness. 76
4-6 Cargo: 50% threshold time to first attempt by attack cost and backup security effectiveness. 77

5-1 Cumulative entries and services in a queue having stochastic entry and service times 86
5-2 Cumulative entries and services in a fluid queue with piecewise constant entry rates 90
5-3 Possible fluid system evolutions when the number of servers at a queue is decreased 91
5-4 Possible fluid system evolutions when the number of servers at a queue is increased 92
5-5 System evolution under greedy server allocations 94

5-6 System evolution under optimal server allocations 96
5-7 Diagram of Boston's Logan International Airport 104
5-8 Expected and actual server allocations to Terminal C under the deterministic fluid model with zero switching times 109
5-9 Expected and actual server allocations to Terminal C under the deterministic fluid model with non-zero switching times 110
5-10 Difference between actual dynamic allocation and original schedule allocation to Terminal C in the stochastically disrupted entry model 114
5-11 Entry rates to Checkpoints C2 and C3 129

B-1 Logistic Regression of Number of Servers Switched, $\theta = 0$ 141
B-2 Logistic Regression of Number of Servers Switched, $\theta = 5$ 143
B-3 Logistic Regression of Number of Servers Switched, $\theta = 10$ 145
B-4 Logistic Regression of Number of Servers Switched, $\theta = 15$ 147
B-5 Logistic Regression of Number of Servers Switched, $\theta = 30$ 149

List of Tables

2.1 Number and percentage of terrorist attacks and fatalities, by attack category 24
2.2 United States population, by year. *Source: U.S. Bureau of the Census [132].* 24
2.3 Total air journeys and average number of journeys per person, by year. *Source: U.S. Air Transport Association [6].* 25
2.4 Time spent per day in each location category *Source: EPA Time Use Survey [154].* 26
2.5 Hourly risk of death in a terrorist attack, by location category 27

4.1 Threshold probability of attempt per flight, r^*, and threshold time for three explosives detection policies. 72

5.1 Sample solutions under fixed, schedule and dynamic allocations of servers . . 87
5.2 Counterexample showing that a greedy allocation, minimizing only the current period's waiting time, is not optimal 95
5.3 Customer entry rates to Terminals C and E at Boston Logan International Airport, January 18, 2005 105
5.4 Average waiting times under fixed and schedule allocations for the deterministic fluid model 106
5.5 Schedule allocations under the deterministic fluid model, $\theta = 0$ or 5 minutes 107
5.6 An example of why switching might increase when switching times are high . 108
5.7 Average waiting time (minutes per customer) under fixed, original and new schedule allocations for the deterministic entry disruption model 111
5.8 Average waiting time under fixed, original and dynamic allocations for stochastic weather disruptions, $p = 1/15$ 112
5.9 Average waiting time under fixed, original and dynamic allocations for stochastic weather disruptions, $p = 1/7$ 115
5.10 Average waiting time under fixed, original and dynamic allocations for randomized entry rates 115
5.11 Average waiting time under fixed, original, hybrid dynamic and nearest neighbor dynamic allocations for the stochastic service times model 116
5.12 Coefficients from the logistic regression of change in server allocation ($n_C - N_C$) on system state characteristics 119
5.13 Performance of the Logistic Regression server allocation model on the training set 119

5.14 Performance of the Logistic Regression server allocation model on a test data set 120
5.15 Average waiting times under the maximum service rate heuristic 122
5.16 Average waiting time in the stochastic entry pattern disruption model when the number of servers is reduced. 124
5.17 Average waiting time in the stochastic service rate model when the number of servers is reduced. 125
5.18 Average waiting time in the stochastic service rate model applied to Checkpoints B2 and B5, for $N = 4$ and $N = 5$. 126
5.19 Average waiting times for the deterministic fluid model schedule allocation applied to Checkpoints C2 and C3. 128

A.1 Fatal Terrorist Attacks Against Americans, 1968-2001 135

Chapter 1

Introduction

Within a period of little more than an hour on September 11, 2001, an unprecedented series of terrorist attacks occurred on American soil. Four commercial passenger planes were taken over by terrorists, three of which were crashed into the Pentagon outside Washington D.C. and the twin towers of the World Trade Center in New York City (precipitating their collapse). It is believed the passengers on the fourth plane prevented an attack on a major target in Washington by taking control of the plane themselves and crashing it into an isolated field in Pennsylvania. Nearly 3000 people died that day.

The United States was not unfamiliar with attacks on aviation. Prior to September 11, nearly 300 American civilians had been killed in aviation related terrorist attacks, most of whom were killed in the 1988 bombing of Pan Am Flight 103 over Scotland. However, the September 11 attacks were significant in their magnitude and in having taken place within the United States. They revealed major weaknesses in the aviation security system and prompted an immediate response. The Aviation and Transportation Security Act (ATSA) [133], passed two months after the attacks, established the Transportation Security Administration (TSA) that would, among other things, be responsible for the screening by a Federal employee of "all passengers and property ... carried aboard a passenger aircraft" (Prior to September 11, most airport security personnel worked for private contractors hired by the airlines). ATSA also required all airports to obtain and use Explosive Detection Systems (EDS) to screen checked luggage by the end of 2002 (Prior to 9/11, only a fraction of checked bags were ever screened). The 2002 Homeland Security Act established the Department of Homeland Security (DHS) to oversee the TSA and other security-related administrations. DHS continues to enact new aviation security regulations (e.g., allowing trained pilots to carry handguns and prohibiting cigarette lighters in carry-on bags [16, 77]) and to examine future threats such as those posed by shoulder-fired missiles [122] and threats to air cargo, general aviation and airport perimeters [147].

While most would agree that aviation security as a whole has improved since the attacks, public debate continues about the value of existing measures and about remaining vulnerabilities. The airline industry, for instance, has suffered financially since the attacks, a phenomenon it attributes partially to a declining economy, but also to the expenses (direct and indirect) of additional security requirements. The Air Transport Association (ATA) es-

timates that the cost of aviation security policy in the first year after the attacks was roughly $4 billion. In addition to this, ATA member airlines cut 100,000 jobs, ground nearly 300 aircraft, eliminated routes and scaled back flight frequencies. Some airlines were also forced to seek bankruptcy protection. The Air Transportation Safety and System Stabilization Act, enacted within weeks of September 11, granted $5 billion in funds directly to the airlines as compensation and also made $10 billion available for loans [7]. Nonetheless, the airlines feel that the burden of aviation security costs, transferred to them in the form of air carrier and passenger taxes, has greatly exceeded this government support, and many are critical of certain security measures that could impose further financial hardship [59, 106]. Former CEO of Delta Air Lines, Leo Mullin, argued, "[T]he industry's ability to address the current crisis is seriously limited by the staggeringly high costs of well-intended post-9/11 actions by the government related to security" [106].

There is also concern that the TSA has implemented its security measures somewhat haphazardly. Terrorism expert Brian Jenkins of the RAND Corporation suggested that the TSA has been "bounced around in a reactive mode" and must "outline a broad strategy" for instituting security measures effectively [54]. Although the TSA recognizes its objective to be providing "tight and effective security in a manner that avoids waste and ensures cost-effective use of taxpayer dollars" (statement of Inspector General Kenneth M. Mead [103]), arriving at such decisions to the satisfaction of all stakeholders can be difficult.

Operations research can help to clarify these often political issues and guide decisions in an objective, analytical manner through the use of mathematical models. Indeed, the National Commission on Terrorist Attacks Upon the United States (the 9/11 Commission) essentially called for increased use of operations research analysis in aviation security policy when it recommended that "the U.S. government should identify and evaluate the transportation assets that need to be protected, set risk-based priorities for defending them, select the most practical and cost-effective ways of doing so, and then develop a plan, budget, and funding to implement the effort" [107]. Operations research can contribute to such a systematic approach by determining the conditions under which a policy may or may not be effective, the level of risk at which a proposed measure becomes cost-effective, and the most efficient implementation of such policies.

It is in these areas that this thesis hopes to illustrate the role that operations research analysis can play in guiding aviation security policy. Some work has already been done in this area, in optimizing the application of security measures to different classes of passengers [68, 69, 71, 100, 156, 157], assessing the performance of multi-tiered security processes [70, 80, 81], evaluating the cost-effectiveness of certain policies [33], and in pointing out potential weaknesses in proposed measures [20, 28]. Obviously, we cannot purport to address in this thesis all issues in aviation security. Rather, we select a few prominent problems arising in the years since September 11, 2001, namely whether securing aviation deserves priority over other potential targets; whether passenger pre-screening systems to identify potentially high-risk passengers are useful; whether explosives detection policies for cargo, airmail and checked luggage carried on passenger aircraft are cost-effective; and whether dynamic assignment of security screeners at various airport security checkpoints can yield major gains in efficiency.

We create mathematical models that try to describe these problems and try to provide guidance in their solution.

We use the next chapter to explore whether or not the recent focus on aviation is a disproportionate response to an isolated event. The 9/11 Commission voiced its concern that United States homeland security policy was focused too heavily on "fighting the last war" [66], by hardening aviation at the expense of other vulnerable homeland targets. Since the attacks, the Department of Homeland Security has spent nearly $15 billion on aviation security compared to only $250 million on transit security [89], and attacks such as the March 2004 Madrid train bombings and the July 2005 subway/bus bombings in London have heightened fears that terrorists may have shifted their focus to ground targets. On the other hand, the Secretary of the Department of Homeland Security, Michael Chertoff, has defended recently this emphasis on aviation, arguing that "a commercial airliner has the capacity to kill 3,000 people. A bomb in a subway car may kill 30 people" [89].

Here we try to contribute to the debate on whether the post-9/11 attention on aviation security is excessive. We conduct a historical study of fatal terrorist attacks against Americans prior to September 11, 2001. The analysis demonstrates that the 9/11 attacks were more a continuation of a terrorist fascination with aviation than a major departure from the existing pattern. As we show, 38% of Americans killed by terrorism over the period from 1968 to September 10, 2001 were victims of aviation-related attacks. In contrast, the average American civilian spends only 0.1% of his time in an airplane or at an airport. From these numbers, we can compute that the hourly risk of being killed in a terrorist attack during an air journey was over 600 times as high as the risk per hour on the ground. And, again, these data pertain exclusively to the period prior to 9/11. Events since September 11 suggest that the threat to aviation continues.

In subsequent chapters, therefore, we explore the efficacy and cost-effectiveness of several security policies adopted or considered since 9/11.

One controversial measure has been the passenger pre-screening system, or profiling system, commonly referred to as CAPPS (Computer-Assisted Passenger Pre-Screening System). Pre-screening systems currently use passenger data collected by the airlines at the time a reservation is made to identify passengers who could potentially pose a higher risk of terrorism. Most passengers undergo only a primary level of screening at the airport security checkpoint, but those passengers selected by the pre-screening system are then subjected to a secondary level of screening, such as a thorough hand-search or explosives trace detection of their carry-on bags.

Supporters of these systems have emphasized the operational improvements such systems permit, because security resources are focused on the individuals who are believed most dangerous, allowing other passengers to pass more easily through the checkpoint. But other people wonder if these assessments about which passengers are dangerous are accurate, and still others wonder whether terrorists can deduce enough about how the systems work to evade them easily. For example, [28] pointed out the possibility that terrorists might probe the pre-screening system. They might send members of their group on trial flights in which they pass through the security checkpoint to ascertain which members are deemed high-

risk and which are deemed low-risk by it. The low-risk members could then be used in an attack and receive only the primary level of screening, avoiding the more stringent scrutiny a high-risk member would receive.

In Chapter 3, we evaluate passenger pre-screening systems, developing a probabilistic model for the likelihood that the terrorists attempt and succeed at a plot to carry weapons on-board the aircraft. This probability is a function of the profiling system's ability to identify potential terrorists and the effectiveness of both the primary and secondary levels of screening. The model considers as well elements of terrorist behavior: we evaluate how this probability changes if terrorists probe the system prior to attacking, or if they can give up on an attack if their chance of success is too low.

We find that while the ability of the profiling system to identify high-risk individuals is often the focus of debate, perhaps of greater importance is the effectiveness of the screening received by both low- and high-risk passengers. Without effective screening, the ability to identify terrorists to receive such screening is of limited value. We also find that the terrorists' ability to probe the system and identify low-risk members amongst themselves can sometimes decrease the likelihood that a successful attack takes place. If the terrorists find by probing that the profiling system is more effective than they expected, they could be sufficiently discouraged to abandon the plot.

Many of the parameters used in the model are unknown, limiting our ability to draw a final conclusion about the effectiveness of profiling systems. However, the contribution of this chapter is to provide a mathematical model that allows policy-makers to assess profiling systems under ranges of reasonable parameter values and to determine their sensitivity to assumptions on terrorist behavior. The model also highlights the importance of key parameters, in addition to the effectiveness of the profiling system, that have been less prominent in the public discussion.

Of regular debate is also the issue of determining under what conditions a particular security measure would be cost-effective. According to Chertoff, we must "maximize our security, but not security at any price" [87], and mathematical models can be used to evaluate when a security measure is worth its price.

Chapter 4 focuses on this cost perspective in the context of three measures to secure the cargo hold of passenger aircraft. The first is Positive Passenger-Bag Match (PPBM), which prevents a piece of checked luggage from remaining on an aircraft if its accompanying passenger fails to board. The second measure we explore is the removal of United States Postal Service airmail packages weighing more than one pound from passenger aircraft, as compared to a proposed policy to screen such packages using bomb-sniffing dogs. The third measure to be considered is a parallel measure for cargo carried on passenger aircraft: given that airmail packages have been removed from passenger aircraft, are we not also obligated to remove cargo?

We use a cost-benefit model to calculate threshold values for the likelihood of an attack beyond which the benefits of a policy exceed the costs. Although the three security measures studied here are similar in their goal of protecting cargo holds, their disparate costs result in quite different thresholds. Because bag-matching is a relatively inexpensive policy to

maintain, it is cost-effective even if the likelihood of an attack on an unmatched bag is quite small. By contrast, removing cargo would cost the airlines billions of dollars in lost revenue and, thus, would not be cost-effective unless an attack were imminent. While the cost-effectiveness of removing airmail from aircraft is somewhat inconclusive, the decision to remove airmail does not automatically set a precedent for removing cargo because the two measures have substantially different decision thresholds.

Although such an analysis is limited by our inability to estimate the likelihood of an attack, the use of a decision threshold in this model permits us to consider only whether the likelihood is higher or lower than this threshold. This mathematical framework can then be used by policy-makers to determine whether society should be willing to incur the cost of a particular security measure.

Once a security measure has been adopted, it is also important to ensure that the policy be implemented as efficiently as possible. When the Aviation and Transportation Security Act mandated the use of Federal employees to screen passengers and baggage at airports, the TSA was faced with the task of hiring as many as 50,000 new employees [54]. Since then, however, limits have been placed on the number of screeners that can be employed at airports, and a recent Senate bill called for a 12% reduction in spending on airport screeners to help fund other security initiatives [88].

Given this pressure to reduce the number of security employees at airports, we examine in Chapter 5 the efficient distribution of screeners to security checkpoint queues within an airport: if an airport has a fixed number of screeners and multiple security checkpoints, how should the screeners be assigned to the checkpoints at different times of day to minimize the time passengers spend waiting in line?

We consider both a pre-determined staffing allocation and a dynamic allocation. The pre-determined schedule specifies, at the start of the day, the allocation of screeners to checkpoints for each time period. Any changes in the allocation to accommodate varying passenger arrival rates to the queues are prescribed in advance and cannot be modified later in the day if, for instance, one queue grows unexpectedly longer than another. A dynamic allocation, however, allows such modifications and can deviate from the original staffing schedule by switching screeners from a shorter queue elsewhere in the airport to the longer queue. Here, we use dynamic programming techniques to formalize when such switches should occur and to assess the benefit of dynamically allocating employees.

Using security checkpoint data from Logan Airport in Boston, we compare the average waiting time spent by an airline passenger in the security checkpoint queue under these two types of allocations. When the only uncertainty in the system is in the time each individual passenger spends being screened, we find that the ability to deviate dynamically from a pre-determined schedule typically yields little additional reduction in the average waiting time of passengers. However, when stochastic impulses affect the system on an aggregate level (such as when the true arrival rate of passengers to the checkpoint deviates significantly from the expected arrival rate), then being able to modify the schedule dynamically in response to these impulses can be beneficial.

Chapter 6 synthesizes the outcomes of our various investigations. We conclude that

different issues in aviation security are amenable to mathematical analysis to varying degrees. The key obstacle facing such analyses is the absence of information. Without knowing certain parameter values (such as the effectiveness of a profiling system at identifying a potential terrorist, or the likelihood of a particular type of attack occurring), and without having a clear understanding of terrorist behavior, it is difficult to arrive at concrete conclusions as to the ability of a policy to thwart terrorism or the expected net value of implementing the policy. Nonetheless, the models presented here serve as tools that can be used to guide policy-makers. Even without knowing the exact value of a given parameter, one might still be able to consider a reasonable range of values and determine an appropriate decision within this range, or explore how the decision changes at the extremes of this range. Moreover, mathematical models can often clarify the reasoning behind qualitative statements about aviation security, and point out the limitations in such reasoning. Problems as important as those arising in Homeland Security could benefit from a wider use of such modeling.

Chapter 2

The Relative Risk of Aviation Terrorism

The nearly simultaneous hijackings of four planes in the 9/11 attacks prompted many new security requirements to prevent future attacks on American soil. It is estimated that the Federal government, through the newly formed Department of Homeland Security and its Transportation Security Administration, has spent about $15 billion directly on aviation security systems since the September 11 attacks (about 90% of the TSA's budget) compared to roughly $250 million for transit security (See, e.g., [74, 89, 128, 129, 135]). This fiscal emphasis on aviation security raises the question of whether the United States is fighting the "last war" (as suggested by [161] and the September 11 Commission [66]), by focusing a large portion of Homeland Security monies on aviation security while leaving other targets (such as ports, power plants and rail systems) as "disaster[s] waiting to happen" (Senator Joseph Biden, cited in [86]). They argue that it is unlikely that terrorists will choose aviation as a future target when so many potentially easier targets exist, so why focus mainly on aviation security, to the detriment of these other targets?

One possible answer is that the new measures have not done enough to render aviation a less attractive target. Reports of the Office of the Inspector General for the Department of Homeland Security suggest that even basic passenger screening is ineffective against certain types of weapons (in particular, explosives) [136, 137, 138]. Another response could be that attacks against aviation cause farther-reaching consequences than attacks involving ground targets. As predicted Gerald L. Dillingham, then of the U.S. General Accounting Office (now known as the Government Accountability Office), in 2000, "Protecting [the air transportation] system demands a high level of vigilance because a single lapse in aviation security can result in hundreds of deaths, destroy equipment worth hundreds of millions of dollars, and have immeasurable negative impacts on the economy and the public's confidence in air travel" [146]. A third justification for the focus on aviation security could be that terrorists

are fascinated by aviation, perhaps partly due to a combination of the above two reasons, and would prefer to attack even a hardened aviation target than another easier but less intriguing target. According to Brian Jenkins, a terrorism expert with the RAND Corporation, terrorists "don't abandon their old playbook. We know that our terrorist opponents continue to be interested in domestic aviation" [54]. This opinion is shared by government officials, including former TSA administrator David Stone, who say that intelligence suggests that al Qaeda and other terrorist groups continue to target aircraft and that aviation continues to pose the greatest risk of terrorism [74, 86, 129].

Indeed, we find in this chapter that prior to September 11, an American civilian was more than 600 times as likely to be killed in a terrorist attack on aviation, per hour, as in an attack on any other activity. We arrive at this conclusion via a relative risk statistic that compares the ratio of hourly risk of death in attacks on aviation, (the fraction of terrorism deaths occurring in aviation over the fraction of time spent in aviation), to that in other types of terrorist attacks. In the next section, we introduce relative risk statistics in a general context. Then, in Section 2.2, we discuss the history of terrorism against Americans, and group terrorist attacks into categories according to where they occurred. Section 2.3 estimates the fraction of time an individual spends in each of the activity categories of Section 2.2.2. These results are used in Section 2.4 to determine the relative risk posed by aviation terrorism. We close in Section 2.5 by discussing some of the recent threats against aviation and argue that the use of historical data can still help assess the threat currently facing aviation security.

2.1 A relative risk factor

Relative risk is used to determine how much more likely an undesirable event is to occur between members of two different populations. Commonly denoted RR, it is simply the ratio of the fraction of Group A members experiencing the event to the fraction of Group B members experiencing the event:

$$RR = \frac{P(\text{Event occurs}|\text{Member of Group A})}{P(\text{Event occurs}|\text{Member of Group B})}. \tag{2.1}$$

If RR is equal to one, then members in Groups A and B are equally likely to experience the event. If $RR > 1$ then the members of Group A are RR times as likely to experience the event than members of Group B. If $RR < 1$, then members of Group B are $1/RR$ times as likely to experience the event than members of Group A (or the members of Group A experience a $1 - RR$ reduction in risk, relative to members of Group B).

We would like to use a relative risk statistic to determine how much more likely an American civilian is to be killed in an act of aviation terrorism, per hour spent in aviation, than in another type of terrorism, per hour spent in all other activities. To do this, we will need a tally of terrorist events against American civilians, categorized by where they occurred, and an estimate of the amount of time Americans spend in aviation- and non-

aviation-related activities. These are discussed in the following two sections.

2.2 Terrorism against Americans, 1968-2001

2.2.1 Definition of terrorism

The United States Federal Bureau of Investigation defines terrorism to be a "violent act or an act dangerous to human life, in violation of the criminal laws of the United States, or of any state, to intimidate or coerce a government, the civilian population, or any segment thereof, in furtherance of political or social objectives" and divides such activity into three categories: a *terrorist incident* that occurred and is attributable to a known terrorist group, a *suspected terrorist incident* that is believed to have been committed by a terrorist group but that has not been fully solved (such as the still unsolved 1975 LaGuardia Airport bombing) and a *prevented terrorist incident* [145]. The FBI also defines the so-called "modern era of terrorism" to begin in the late 1960's, and it is in 1968 that the U.S. government began keeping formal records of terrorist events against Americans. We use this as the starting point for our study.

We consider domestic and international terrorist incidents and suspected terrorist incidents in which American civilians were killed from the period of January 1, 1968 through September 10, 2001 (we exclude the terrorist attacks of September 11, 2001 because that single data point would have been responsible for 80% of all terrorist fatalities against Americans). We exclude attacks that were targeted specifically against military facilities (such as the 1983 attack on a Marine barracks in Beirut, Lebanon) because such an attack could be considered an act of war. We do, however, include attacks on military personnel outside of military bases because such attacks are similar to those experienced by non-military personnel. Using publicly available sources [1, 2, 17, 26, 101, 125, 145, 151, 152, 153], many of which are official U.S. reports on terrorism against Americans, we have created a list of terrorist events in which American civilians were killed, given in Appendix A. In total, there were 179 fatal attacks, killing 740 American civilians, ranging from small attacks such as kidnappings of a single person to large attacks such as the bombing of Pan Am Flight 103, killing 189 Americans, and the Oklahoma City bombing, which killed 168.

2.2.2 Classification of terrorist attacks

In many of the sources, the location and context of each attack were given, allowing us to attribute the attack to one of five categories: Aviation, Work, Home, Leisure, Other Travel. In other sources, such details were not given. Twenty-seven attacks (many of them kidnappings where it is unknown or unstated from where the victims were kidnapped) had no clear category given, and we assigned a likely category to these attacks based on similar attacks in which the location was stated. Table 2.1 shows the number and percentage of attacks and the number and percentage of fatalities by attack category. We see that while the number of attacks on aviation (24 out of 179) is small compared to the other categories, it

Category	Number of Attacks	Percentage of of Attacks	Number of Deaths	Percentage of Deaths	Average Fatalities per Event
Aviation	24	13.4%	294	39.7%	12.2
Work	40	22.3	260	35.1	6.5
Leisure	38	21.2	80	10.8	2.1
Other Travel	40	22.3	63	8.5	1.6
Home	37	20.7	43	5.8	1.2

Table 2.1: Number and percentage of terrorist attacks and fatalities, by attack category

Year	Total U.S. Population (millions)
1970	203
1975	216
1980	231
1985	238
1990	253
1995	263
2000	286
Average	240

Table 2.2: United States population, by year. *Source: U.S. Bureau of the Census [132].*

accounts for the greatest share of terrorist fatalities (39.7%) and the highest average number of fatalities per attack, 12.2, while attacks on the workplace, the next highest, average around 6.5 fatalities. The large share of fatalities attributable to aviation is especially striking when one considers how little time the average American civilian spends in aviation related activities compared to time spent, say, at home or at work.

2.3 Fraction of time spent in aviation

The historical likelihood of being killed in a terrorist attack per hour is the total number of terrorism fatalities divided by the total number of person-hours over our time horizon. To find the total person-hours over the period of 1968-2001, we look at the United States population over this period at five-year marks along the interval (shown in Table 2.2), estimate a time average population of 240 million, and multiply this by the total number of hours in the period from 1968 to 2001. This yields a value of roughly $7.15 \text{x} 10^{13}$ person-hours. So the total risk per hour to an American civilian of being killed in a terrorist attack is the number of fatalities, 740, divided by the number of person-hours, $7.15 \text{x} 10^{13}$, or $1.03 \text{x} 10^{-11}$ (equivalent to waiting an average of 11 million years before succumbing to such an attack). To find a similar value specific to aviation attacks, we use the fraction of these person-hours spent in

Year	Total Air Journeys on U.S. Airlines (millions)	Average Enplanements per person
1970	154	0.76
1975	189	0.88
1980	273	1.18
1985	357	1.50
1990	424	1.68
1995	499	1.90
2000	611	2.14
Average Ratio (weighted by pop.)		1.46

Table 2.3: Total air journeys and average number of journeys per person, by year. *Source: U.S. Air Transport Association [6].*

aviation. There are two ways at arriving at such a fraction, and we will show that both yield roughly the same value.

2.3.1 Estimation based on passenger enplanements

We can use annual passenger enplanement data from the Air Transport Association to estimate the average number of flights per year that an American takes and from there, estimate the amount of time spent at the airport or onboard the aircraft. Passenger enplanements per year are shown in Table 2.3, as are the ratios of the number of journeys to the population in each year. Over this 34-year period, each American citizen took approximately 1.5 flight legs on domestic flights annually[1].

We need next to translate this average number of flights per year into an average time spent at an airport or on an aircraft. Before September 11, 2001, air passengers arrived roughly one hour prior to takeoff. An average non-stop flight, being approximately 1000 miles long, would have taken about 2.5 hours by jet. However, one-third of passengers require connecting flights, adding an additional hour to the total travel time. Lastly, the typical passenger generally leaves the arrival airport within 20 minutes after disembarking the plane. Thus, the total time per flight is approximately (1 + 2.5 + 1/3*1 + 1/3) = 4.2 hours. If we allow for occasional flight or other delays, then we suppose that the average total time per air journey is approximately five hours[2]. Thus, the total person-hours spent in aviation-related activities from 1968-2001 was roughly 6.1×10^{10} person-hours, or 0.1% of the total person-hours in that time. An American citizen spent roughly one-tenth of one percent of their time at an airport or in an airplane.

[1] While not all passengers on domestic flights are U.S. citizens, and not all Americans travel solely on U.S. domestic flights, we assume that these two effects roughly cancel each other.

[2] While this estimate is based on domestic flights, and we must include international flights as well (which can be significantly longer), this estimate of five hours can still be appropriate when one considers that many international flights are themselves quite short, such as between Chicago and Toronto, or Tokyo and Seoul.

Category	Total Minutes per Day	Fraction of Time per Day
Aviation	1.43	0.001
Work	179.28	0.124
Leisure	119.86	0.083
Other Travel	91.36	0.063
Home	1048.07	0.728

Table 2.4: Total time and fraction of time spent per day in each location category *Source: EPA Time Use Survey [154].*

2.3.2 Estimation based on time-use studies

Another method to determine the fraction of time an American spends in each of the categories given in Section 2.2.2 is to use Time Use Survey data that report the average amount of time spent by survey participants in various locations. We referred to the Environmental Protection Agency National Time Use Survey [154]. This nationwide telephone survey was conducted from September 1992 through October 1994 and published in 1995. From each participating household, one person (either an adult or a child under the age of 18) responded, and a total of 9386 surveys (from 7514 adults and 1872 children) were completed. The survey consisted of participants listing each activity they had engaged in and each location they had visited the day before, as well as an estimate of the time spent in each. We chose this time study over others because it distinguished air travel from other modes of transportation. A summary of the total time spent (in minutes) per day in each location category, as defined in Section 2.2.2, is shown in Table 2.4. Once again, we see that the average American spends about 0.1% of their time either at an airport or on an airplane.

2.4 Death risk per hour

As we saw in Section 2.3, the risk per hour to an American civilian of being killed in a terrorist attack of any type is approximately $1.03\text{x}10^{-11}$. However, if we divide the number of terrorist fatalities in a particular category by the number of person-hours spent in that category (which is the total number of person-hours over 1968-2001 times the corresponding fraction of time from Table 2.4), then we obtain the risk per hour of being killed in an attack on that category. These hourly risks are given in Table 2.5. While the hourly risk of death in an aviation terrorist attack is small in absolute terms, it is greater than that of any of the other categories.

Using Equation (2.1), the relative hourly risk posed by aviation terrorism versus non-aviation terrorism is given by:

$$RR = \frac{\text{Hourly Risk from Aviation}}{\text{Hourly Risk from Non-Aviation}}$$

Category	Hourly Risk
Aviation	**$4.11\text{x}10^{-9}$**
Non-Aviation	**$6.24\text{x}10^{-12}$**
Work	$2.93\text{x}10^{-11}$
Leisure	$1.35\text{x}10^{-11}$
Other Travel	$1.38\text{x}10^{-11}$
Home	$8.26\text{x}10^{-13}$

Table 2.5: Hourly risk of death in a terrorist attack, by location category

$$\begin{aligned} &= \frac{4.11\text{x}10^{-9}}{6.24\text{x}10^{-12}} \\ &= 658.6. \end{aligned} \tag{2.2}$$

An average American civilian is more than 600 times likely to die in a terrorist attack on aviation, per hour, than in all other activities.

2.5 Using the past to predict the future

Clearly, prior to September 11, 2001, aviation attacks caused a disproportionately large fraction of the total American civilian terrorism fatalities, and on 9/11 they more than quadrupled the total number of American terrorism fatalities, via a series of attacks on aviation. Does terrorists' fascination with aviation continue?

More rigorous passenger and baggage screening may have reduced the likelihood of another attack on aviation through both improved detection capabilities as well as deterrence effects. Perhaps this heightened security has caused terrorist groups to shift their focus to different targets, increasing the risk of non-aviation terrorism relative to aviation terrorism. On the other hand, many argue that September 11 marks the beginning of a new era of terrorism against the United States, and terrorists, spurred on by their successful 2001 attacks, might continue to plot even more devastating aviation attacks.

The evidence supports this latter claim. A few months after the 9/11 attacks, Richard Reid attempted to ignite explosives hidden in his shoes on a flight from England to the U.S., and in July 2002, three people were shot and killed by a gunman at Los Angeles International Airport. Even as recently as April 8, 2005, two passengers on a KLM flight from Amsterdam to Mexico (scheduled to fly over U.S. airspace) were discovered to be on the U.S. no-fly list and the flight was sent back to Amsterdam. International events also serve as evidence that the use of aviation as a vehicle for terrorism continues. There was, for instance, the near-miss of shoulder-fired missiles on an Israeli airline over Mombasa, Kenya in 2002 and the successful downing of two passenger planes in Russia in August 2004. A study published by the RAND Corporation examining the cost-effectiveness of various anti-missile technologies argues that because terrorists have the motive and means to carry out an attack using shoulder-fired

missiles and are simply waiting for an opportunity to use them, the threat of such an attack could be high [33].

In addition to these known attempted and successful attacks, the U.S. government has repeatedly mentioned intelligence reports suggesting that al Qaeda's interest in aviation continues even now, almost four years after 9/11. In 2003, officials warned of the possibility of another 9/11-style attack in which terrorist teams might hijack aircraft, and an undercover FBI agent intercepted the sale of a shoulder-fired missile by a British arms dealer who believed he had been interacting with terrorists. In 2004, the Department of Homeland Security suggested that terrorists might try to blind airline pilots with high-power lasers, a possibility made more grim by a series of incidents in which pilots did notice lasers pointed at their aircraft (although it was later believed to have been innocent civilians improperly using laser technology) [49]. An FBI/DHS report released in early 2005 indicates that terrorist groups continue to explore aviation as a means of attack, even testing the system to discover loopholes [74], and while the report focused specifically on general aviation and helicopters, it emphasized that the threat on commercial aviation is ever-present as well. Furthermore, it continues by saying that improvements since September 11 have "reduced, but not eliminated" the threat of future attacks on aviation [86]. Former Transportation Security Administrator David Stone warned of "threat streams" indicating that the greatest risk for future terrorism still lies within the aviation sector [74, 86, 128].

Clearly, aviation terrorism was a threat long before September 11, on and immediately after September 11, and it continues to be a threat. Given that so many resources are being devoted to aviation security, it is worthwhile to consider how those resources are being used and whether they are efficient at achieving their stated goals. The remainder of this thesis attempts to use mathematical modeling to guide the discussion. We begin in the next chapter with the question of how to evaluate the effectiveness of a security measure, in the context of computerized passenger profiling.

Chapter 3

How Effective Might Passenger Profiling Be?

While some aviation security measures are applied equally to all passengers, others are considered so time-consuming that they are restricted to a fraction of air travelers. A key issue in the debate is the question of which passengers should be subjected to special scrutiny at the airport prior to boarding the airplane and how they can be identified. Since 1973, all passengers on American airliners have been required to pass through metal detectors or pass their carry-on baggage through x-ray machines in an attempt to prevent hijackers from carrying weapons aboard the plane. In 1997, at the recommendation of the White House Commission on Aviation Safety and Security (also known as the Gore Commission) [162], a system was introduced to identify potentially dangerous passengers whose checked luggage would be screened for explosives. This system, known as the Computer Assisted Passenger Pre-screening System (CAPPS), used a set of risk-determining criteria to try to identify potentially high-risk passengers whose checked bags needed further investigation, and it also selected for screening the bags of a fraction of presumably low-risk passengers chosen at random. After September 11, however, the policy was extended such that any passenger chosen by CAPPS would now undergo additional personal screening, such as hand-wand inspection, pat-down inspection and/or hand searching of carry-on bags at the security checkpoint or at the gate prior to boarding.

The method for choosing "selectees" has been changing over time, as has its name: first CAPPS, then CAPPS II and now Secure Flight, with each successor trying to find an acceptable balance of collecting government and airline data that could lead to terrorists, while respecting civil liberties. The original CAPPS uses only information obtained by the airlines at the time a ticket is purchased. Commonly cited risk indicators are purchasing one-way tickets or paying in cash. The initial proposed successor to CAPPS, CAPPS II, would have used additional data, such as credit reports, criminal records, travel history, cell phone calls among other personal information [96, 110]. In response to privacy concerns, the TSA modified its description of CAPPS II to reassure the public that only names, addresses and dates of birth would be used to verify passengers' identities and compare them to criminal and intelligence databases [76, 78, 139, 142, 143]. However, due to concerns

that criminal records might be misused to capture non-violent criminals [8, 76, 110, 111], and TSA's inability to collect passenger data from the airlines, CAPPS II was dropped in July 2004. A new profiling system, Secure Flight, was proposed shortly thereafter and is currently in development. This system will be managed entirely by the government, merging federal, non-criminal databases, and will apply only to domestic flights [144, 148, 160]. The Transportation Security Administration (TSA) hopes to have Secure Flight in place by the end of 2005, but in the meantime, the original CAPPS is still in use. The analysis presented in this chapter does not rely on the type of data used by the profiling system, nor do we address the civil liberties issues associated with such a system. (The interested reader can read a Congressional debate of such issues in [149]). Instead, we focus on the quantitative performance of pre-screening systems.

There are divergent opinions regarding the anti-terrorist effectiveness of such passenger pre-screening systems. Secretary of Transportation Norman Y. Mineta described a strong pre-screening system as the "foundation" of aviation security [109]. Supporters of profiling systems feel that they allow a more efficient allocation of security resources. The heightened security after 9/11 forced passengers to arrive earlier at the airport than they were accustomed to doing. This so-called "hassle factor" is believed to have caused a reduction in the demand for air travel shortly after September 11. In 2002, Leo Mullin, then CEO of Delta Air Lines, estimated the cost of this hassle factor to be $3.8 billion industry-wide in 2001-2002 [59]. Although security delays have since diminished, many passengers are still frustrated because they believe it should be obvious they pose no threat and thus should not be subjected to random searches. Donald Carty, former CEO of American Airlines, has said, "[CAPPS II is] simply going to allow us, instead of focusing on the kind of thorough security procedures we're going through, with every single person that goes through...to focus resources on really suspicious targets. And that's the right answer to security" [73]. Former TSA Administrator Admiral James M. Loy described profiling systems as "a valuable tool in holding down passenger wait times by reducing the number of people who undergo secondary screening or who are misidentified as potential terrorists" [140]. Furthermore, there is a widely held belief that the presence of such security measures alone will deter terrorists from even attempting to attack. According to Carty, "With the amount of security that we have in the aviation system today, the likelihood of a terrorist choosing aviation as the venue for future attack is very low" [15].

Others, however, have questioned the protection offered by pre-screening systems. One problem that has received special attention is the ability of terrorist groups to "reverse engineer" the system and thereby thwart it. As Chakrabarti and Strauss [28] have argued, terrorist groups can find out through test flights which of their members are selected by computer for secondary screening and which are not, a process the authors liken to a carnival game. Then in actual missions, group members classified as low-risk could take the lead roles. In consequence, the true effectiveness of the pre-screening system might be far less than hypothesized. Others feel that computerized passenger profiling may not be an effective security tool. Brent Turvey, forensic scientist and criminal profiler argued, "You can't use profiling to predict crimes, only analyze crimes that are in the past" [62], and Barnett [20]

outlines several problems with attempting to use assumed terrorist characteristics to identify future terrorists.

Who is closer to the truth, the optimists or the pessimists? We consider that issue in this chapter, though we cannot provide a definitive answer. Instead, we develop a parameterized probabilistic model of the pre-screening process, one component of which relates to terrorist behavior, and use it to evaluate the probability that a terrorist is able to board a commercial aircraft with the intent to attack it. We investigate which parameters are most important for reducing the probability of a successful terrorist attack. Moreover, by clarifying the tacit assumptions underlying the opposing opinions, the formulation identifies some potential problems with their arguments. In some circumstances, for example, the ability of terrorists to probe the screening system can actually reduce the danger of a terrorist attack. And even if the pre-screening system is extremely good at identifying high-risk individuals, it might only minimally reduce the chance of successful terrorism if the secondary screening process is inadequate. We find that if the profiling system is not as effective as we might hope, or not robust to terrorist behavior, our efforts would be better applied to improving the primary screening imposed on all passengers. However, if the profiling system does a good job at identifying terrorists, then it is imperative that we develop a strong secondary screening process.

In the next section, we describe the airport security checkpoint process and define parameters associated with key components of this process. Section 3.2 discusses, through the use of these parameters, the opposing opinions on pre-screening systems and their underlying assumptions. In Section 3.3, we develop our model, incorporating the terrorists' probing behavior as well as deterrence effects. Section 3.4 describes the methodology used to analyze this model in which none of the parameters are known exactly. We present in Section 3.5 the results of our analysis, including some counterexamples to the conclusions drawn by those factions. We summarize our conclusions and policy recommendations in Section 3.6.

3.1 System and parameter description

When a passenger checks in for a flight, the passenger screening system (hereafter PSS) labels her as either low- or high-risk and indicates this status on her boarding pass. She then proceeds to the security checkpoint where, as shown in Figure 3-1, if she is labeled low-risk and is also not selected at random to receive additional scrutiny, she passes through the metal detector and sends her carry-on bags through the x-ray machine, a process we call *primary screening*. Otherwise, she must pass first through primary screening and then undergo a more thorough search of her belongings and clothing. This entire alternate process, which includes the initial primary screening, is referred to as *secondary screening*.

We define the following parameters:

- C, the *a priori* probability that an actual terrorist is classified as high risk by the PSS.
- r, the percentage of "low-risk" passengers selected at random for secondary screening.

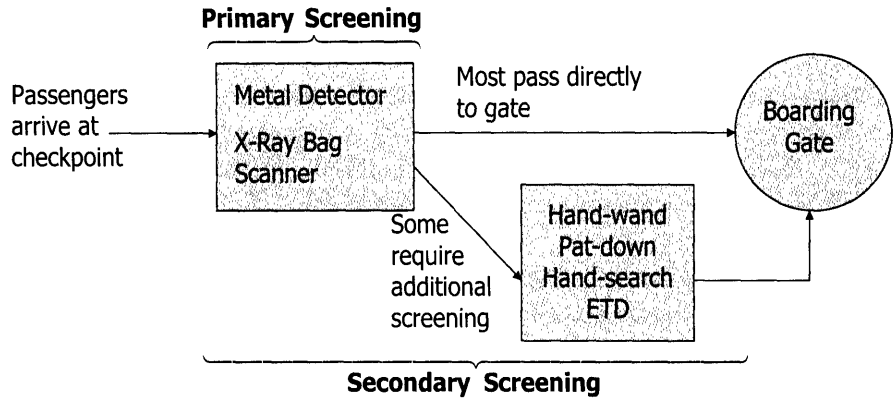

Figure 3-1: Current security checkpoint procedure.

- p_1, the conditional probability that primary screening detects the terrorist's weapons and prevents him from proceeding further, given that he receives only primary screening.
- p_2, the conditional probability that secondary screening detects the terrorist's weapons and prevents him from proceeding further, given that he receives secondary screening. Because secondary screening includes primary screening, we have $p_2 \geq p_1$.
- τ, the terrorist group's *deterrence threshold*: the minimum probability of success required by the group to proceed with the attack.
- n, the number of potential participants in a particular terrorist plot.

We note that C, p_1 and p_2 depend heavily on the particular members of the terrorist plot and on the particular nature of the plot itself. For instance, an x-ray machine might be very good at detecting a loaded gun, but less effective at detecting a small knife and virtually ineffective at detecting plastic explosives. We also assume that these parameters reflect the terrorists' estimate of the values they represent and that their estimates are realistic. Moreover, the value, C, does not denote the overall fraction of passengers that are selected for secondary screening but is instead a conditional probability that a *terrorist* will be selected by the profiling system. ([100] present a model for determining the fraction of passengers to be assigned to different risk classes to minimize the chance of a successful attack, subject to budget constraints on the number of passengers undergoing stringent screening).

The use of a threshold parameter to model deterrence requires some explanation. A vulnerability analysis, as defined in [53], examines the probability that an attempted attack

on a system is successful, a measure of a system's inability to thwart attempts against it as they occur. By contrast, a complete risk analysis would also take into account the likelihood that an attack will be attempted in the first place, considering the frequency with which such attacks occur. Here, because we do not consider the total frequency with which plots occur, we are not performing a complete risk analysis. However, we expand on vulnerability analysis to consider not just the system's ability to detect plots in progress, but also its ability to deter terrorists from attempting their plots. If, as is widely believed (see, e.g., [41]), terrorists consider the costs, value and likelihood of success of an attack, then using a threshold parameter is reasonable. They will attack when the value of an attack (weighted by its chance of success) exceeds its cost (weighted by its chance of failure). If their estimated probability of success is greater than a threshold, τ, then they will attack, otherwise they will not. Low values of τ reflect willingness to attack despite a high risk of failure (for instance, when the cost of a failed attack is relatively low). Higher values of τ indicate risk aversion.

A similar method was used in [11, 12, 150] for drug interdiction. In those studies, law enforcement officers interviewed several drug smugglers to determine each of their individual threshold levels as a function of the probability and legal consequences of getting caught, and then fit a function to represent the average of these thresholds. While there is substantial data that can be used to estimate deterrence effects on drug trafficking as a function of interdiction efforts (in addition to using interview data, drug prices before and after interdiction can be examined, as in [40]), there is little such data available for terrorist behavior, so calibrating a "willingness to attack" function is impossible at this time. Nonetheless, the use of a deterrence threshold in the context of drug smuggling supports our use of a threshold here. We note that although deterring an on-board attack might only lead to mayhem elsewhere, this caveat can be applied to any successful anti-terrorist measure. Indeed, the Office of National Drug Control Policy defined a drug interdiction effort to have a "deterrent effect" if it at least forced the smugglers to use a less attractive route that was riskier or more costly to them [150]. In our model, we use this same notion and focus only on terrorists' ability to commit their original intended attack.

Using these parameters, we can estimate the probability that a terrorist can successfully board an aircraft with weapons. For ease of exposition, we assume that an attack is attempted by a lone terrorist, chosen out of the group of n, who tries to pass through screening. (Two lone terrorists with plastic explosives destroyed two Russian passenger planes in August 2004). We treat the attack as successful if he makes it through the screening process with his weapons and boards the aircraft, thus assuming that once he has boarded the aircraft, his attack will proceed unthwarted. As shown in Figure 3-2,

$$\begin{aligned} P(\text{Terrorist receives primary screening}) &= (1-C)(1-r), \\ P(\text{Terrorist receives secondary screening}) &= C + (1-C)r. \end{aligned}$$

Hence, a successful attack occurs with probability

$$(1-C)(1-r)(1-p_1) + (C + (1-C)r)(1-p_2). \tag{3.1}$$

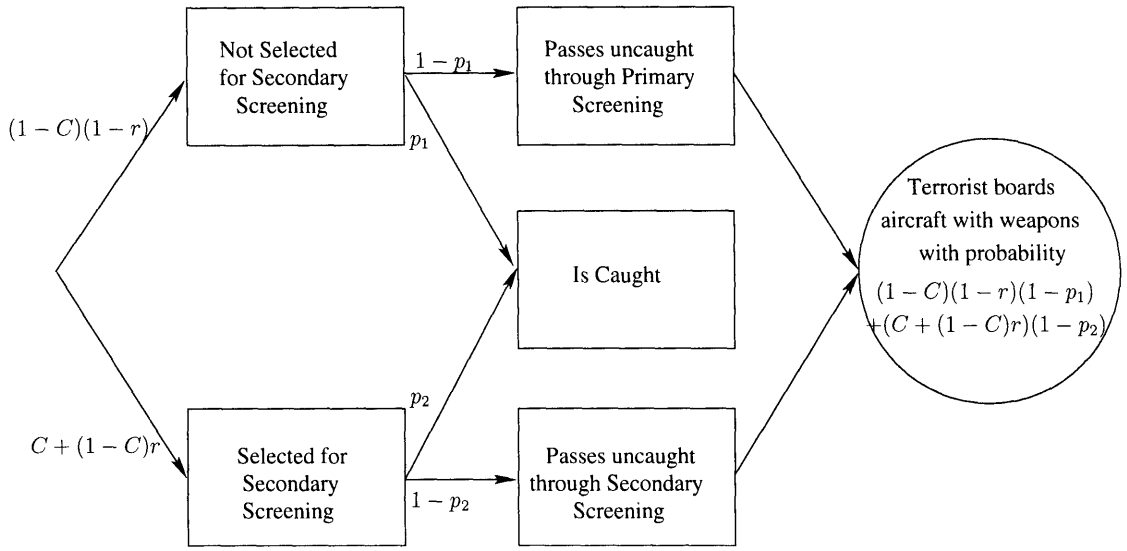

Figure 3-2: The path through airport passenger screening for an individual terrorist carrying weapons

If his calculated chance of success is smaller than τ then he will forego his attack, otherwise he will proceed.

3.2 Opposing viewpoints

As noted, there is much debate about whether a profiling system can substantially improve aviation security. We discuss now the arguments of both proponents and critics in terms of the parameters defined above.

3.2.1 The "right answer" to security?

Supporters of pre-screening systems argue that if a computerized profiling system can identify terrorists, then we need screen only these high-risk individuals. Random searching would not be necessary, and even general passenger screening would be less important, facilitating the flow of passengers through the checkpoints. Furthermore, if this focused security were very effective, terrorists would be deterred from using commercial aviation as their target for attacks.

They assume, given their wording, that the PSS will flag virtually all terrorists as high-risk, or that C is very close to 1. Their belief that screening of low-risk passengers would be expedited by a profiling system indicates an assumption that the PSS will not select too many "innocent" passengers, (i.e., the false positive rate is reasonably small). Furthermore, some believe that terrorists can be deterred by the prospect of stringent searching (i.e. the terrorists have a nonzero deterrence threshold, τ). Under these assumptions, almost all terrorists who attempt to board the plane will receive secondary screening, and the

probability that they attempt and succeed at their attack will be either $1 - p_2$ (if they attempt the attack) or zero (if they are deterred). Under an (implicit) assumption of highly effective secondary screening, the terrorists' probability of success could approach zero even if they are undeterred.

3.2.2 A carnival game?

There are others who feel that profiling introduces weaknesses to the security system. Barnett [20] questions the assumption that C would be very high, citing, among other concerns, the difficulties in using limited information about past terrorists to identify future terrorists, and the possibility of terrorists using innocent-seeming "dupes" who do not realize they are carrying deadly items.

Chakrabarti and Strauss [28] go even further and argue that profiling applied to passenger screening could actually help terrorists improve their chance of success. Terrorist groups can send members on trial flights, without the intent to attack. Because the additional screening of a selectee is of a personal nature, they can ascertain who is considered low-risk by the pre-screening system based on who is not pulled aside for secondary screening. If we assume that passengers' PSS scores are constant during the probing-and-attack cycle, then a terrorist who passes unflagged even once through the checkpoint knows for certain he has a low PSS score. In the real attack, the group could use this low-risk member knowing that he will face additional scrutiny only if chosen at random. Essentially, each terrorist "steps right up" to the security checkpoint to discover which level of screening he is to undergo.

Moreover, they argue that a terrorist group can *always* find such a "low-risk" group member. First, they assume that n, the pool of terrorist probes, is arbitrarily large. Second, they assume that $C < 1$ and that risk ratings for different group members are independent. Thus, even if k terrorists have been flagged by the profiling system, the probability that the $k + 1^{st}$ is selected remains C, suggesting that the profiling system's ability to detect ties between members of a same terrorist group is poor. More specifically, they assume each passengers' score to be selected independently from a Gaussian distribution, with the terrorists having a higher mean score than innocent passengers. The profiling system selects, for each flight, a given percentage of passengers having the highest profiling scores to receive secondary screening. As such, in their model, a passenger might be considered high-risk on one flight but low-risk on the next.

The independence assumption implies the probing process is Bernoulli, and, if the group keeps sending probes, eventually one will be classified as "low-risk". At the time of the attack, C would be zero for such a known "low-risk" passenger, and the probability that he would succeed would be as in Equation (3.1) and Figure 3-2, with $C = 0$:

$$(1 - r)(1 - p_1) + r(1 - p_2).$$

While the authors do not discuss deterrence, they tacitly assume that

$$\tau < (1 - r)(1 - p_1) + r(1 - p_2),$$

so the terrorists will not be deterred. Because the chance of receiving secondary screening in this situation is simply r, they conclude that using a PSS would do no better than random screening alone.

3.3 A general model

How crucial are the assumptions presented above to the final conclusions drawn by the optimists and pessimists? We will develop in this section a general model, capable of accommodating the various sets of assumptions, to evaluate a profiling system's ability to reduce the likelihood of a successful attack. We begin by stating our own assumptions.

3.3.1 How terrorists might probe the system

We will use the same parameters and security checkpoint procedure described in Section 3.1, but we must formalize our assumptions about how the pre-screening system selects passengers and the terrorists' probing process.

The National Research Council's Committee on Science and Technology for Countering Terrorism [37] indicates that profiling systems select passengers for additional screening according to an absolute risk score. Passengers are rated "Green", "Yellow" and "Red" for low-risk, high-risk and no-fly status (for known terrorists), respectively. It is estimated that in a new profiling system, approximately 5-8% of all passengers would be labeled as Yellow and require additional screening (compared to 15% under the original CAPPS), and that fewer than 2% would not be allowed to fly [8, 76, 78, 123]. We assume this status does not change from flight to flight. Thus, if a passenger passes once through the security checkpoint without being selected, he is certain to have a low-risk classification on subsequent flights. On the other hand, selection for secondary screening does not always imply high-risk status, as a fraction r of low-risk passengers are selected at random for this screening. Though this is different from the "curved grading" used in [28], where one might be among the highest scoring $x\%$ of passengers on one flight but not another, this absolute scoring seems more logical, aligns well with the available information, and simplifies our analysis.

We model the attack process as a sequence of decisions made by the terrorists. At each stage, they compare their estimated chance of success from the current stage onward to their deterrence threshold τ. If ever it falls below τ, they give up and do not proceed to the next stage. Note that taking $\tau = 0$ is equivalent to a model where terrorists are assumed never to give up.

First, the group decides, based on an initial estimate of their likelihood of success, whether to take the first step towards an attack by probing the system, or to give up at the outset (if their initial estimated chance of success is lower than τ). Assuming they decide to probe, all n members[1] are sent, in sequence, on probing flights where they each pass once through the

[1] In our model, the group never decides to give up in the middle of the probing process. Either no member is sent and they give up outright, or all are sent, at least until a low-profile member is found. One can imagine a different model where, in the middle of the probing process, the terrorists use the results of previous trials

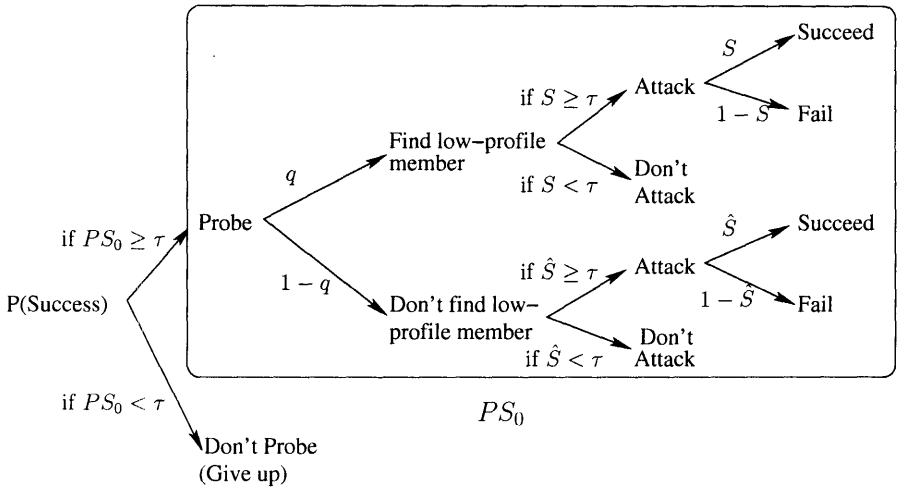

Figure 3-3: A decision tree describing the probability of a successful attack. The shaded portion represents the value PS_0, the probability of success once the terrorists have decided to probe. The decision of whether or not to probe is made based on comparing PS_0 to the value τ.

security checkpoint and notice whether they are pulled aside for additional screening or not. After the flights, the group can use the results to update their estimate of C, the chance of being selected by the PSS. Thus, if k members were pulled aside for additional screening, the terrorists might infer that the probability of selection for the $k+1^{st}$ is higher, relaxing the independence assumption used by Chakrabarti and Strauss. (Holding the value of C constant is equivalent to treating the terrorists as independent).

After all of the probing flights, the group decides whether or not to commit the attack. If a low-profile member was found during the probing flights, then he is sent on the attack. Otherwise, if no low-profile member was found, the group must decide whether to attempt the attack with one of the previously flagged members (if the chance of success using a previous selectee is sufficiently high) or to give up without attacking.

3.3.2 Modeling the probability of a successful attack

We now formalize this two-stage model, for which the decision tree is shown in Figure 3-3. As shown in the shaded portion, if the terrorists have decided to probe, then in the best case, they will find at least one group member who is not selected for secondary screening during his probing run. We assume that this happens with probability q, which depends on C and on the updating scheme the leader might use, to be discussed in Section 3.3.3.

to decide whether to continue probing or to give up immediately. Because we assume that the probing flights contribute negligible cost, the terrorists will send all n members to probe if they feel they have a reasonable chance of finding at least one low-profile member.

Concluding that he has a low PSS score, they can send him on the attack, where only if flagged at random will he be subjected to additional screening. In this case, the attack's probability of success, S, is

$$S = (1-r)(1-p_1) + r(1-p_2). \tag{3.2}$$

If $S \geq \tau$, then the group will decide to attack using this low-profile member. If $S < \tau$, then even a low-profile terrorist does not guarantee a sufficiently high probability of success, and the group will not attack.

In the worst case, occurring with probability $1-q$, all n members will be selected for secondary screening (but they do not know which among them have been selected by the PSS and which at random). Despite this, they might feel that they have still a reasonable probability of success using one of these possibly high-profile group members in the attack. The group leader can update C, the estimated probability of a group member being flagged by the PSS, to a higher value, $\hat{C}$, based on the results of these probings, and use this updated value in Equation (3.1) to estimate the chance of success using a previously flagged member in the attack. We call this value $\hat{S}$:

$$\hat{S} = (1-\hat{C})(1-r)(1-p_1) + \left(\hat{C} + (1-\hat{C})r\right)(1-p_2). \tag{3.3}$$

If $\hat{S} < \tau$, then the terrorists will not attack. Otherwise, if $\hat{S} \geq \tau$, then even if all members are selected for additional screening during the probing runs, they will still proceed with the attack, and succeed with probability $\hat{S}$. We note that our assumption that $p_2 \geq p_1$ implies that $S \geq \hat{S}$.

The above values give the estimated probability of success after the terrorists have decided to probe the system. The question remains of whether the terrorists should probe in the first place, or just give up at the outset. For instance, if $S \geq \tau$ but $\hat{S} < \tau$, the terrorists will attack only if they find a low-profile group member during the trial flights. If the probability, q, of them doing so is very small, then they may decide probing is not worth the effort. We let PS_0 be the total probability of success anticipating n probing runs:

$$PS_0 = \begin{cases} qS + (1-q)\hat{S} & : \quad S \geq \hat{S} \geq \tau \\ qS & : \quad S \geq \tau, \hat{S} < \tau \\ 0 & : \quad \hat{S} \leq S < \tau. \end{cases} \tag{3.4}$$

To decide whether or not to probe, PS_0 is compared to τ. If $PS_0 \geq \tau$, then the terrorists proceed with the probing flights, otherwise the terrorists give up immediately. In the case where $S \geq \hat{S} \geq \tau$, the terrorists will always attack, even if all group members are flagged in the probing flights. However, the group still sends its n members on probing runs in case they can find a low-profile member for whom the probability of success would be greater. $S \geq \tau, \hat{S} < \tau$ represents the case where the probability of success is acceptable if a low-profile member is found, but not acceptable when only a previously flagged member is available. In this case, the n probes will be sent only if there is a reasonably high chance, q, of finding a

low-profile member ($qS \geq \tau$), and the attack is attempted only if a low-profile group member is found during the course of the probings. Lastly, if we have $S < \tau$, then $\hat{S}$ is also smaller than τ and $PS_0 = 0$; no probes are sent and the attack is not attempted.

The overall probability that an attack is **A**ttempted and **S**uccessful, $P(AS)$, is therefore

$$\begin{aligned} P(AS) &= \begin{cases} qS + (1-q)\hat{S} & : \quad PS_0 \geq \tau, S \geq \tau, \hat{S} \geq \tau \\ qS & : \quad PS_0 \geq \tau, S \geq \tau, \hat{S} < \tau \\ 0 & : \quad PS_0 < \tau \end{cases} \\ &= \begin{cases} qS + (1-q)\hat{S} & : \quad S \geq \tau, \hat{S} \geq \tau \\ qS & : \quad qS \geq \tau, \hat{S} < \tau \\ 0 & : \quad qS < \tau, \hat{S} < \tau. \end{cases} \end{aligned} \tag{3.5}$$

The task still remains of finding values for q, the likelihood of finding a low-profile member, and $\hat{C}$ (used for computing $\hat{S}$), the updated chance that a terrorist will be selected by the PSS at the time of the attack, *given* that he and all of the other group members were pulled aside for additional screening during the probing runs. These values depend on how we assume the PSS detects links between members of a same terrorist group, as we will discuss in the next section.

3.3.3 Updating conditional probabilities of selection

As mentioned earlier, if all probing runs fail (meaning all n group members are selected either by the profiling system or at random during the trials), the group leader might adjust the estimate of C based on this information to some new value $\hat{C}$, the conditional probability that a member has a high PSS score given that all group members were selectees in the trial flights. His method for doing this depends on how the profiling system detects associations between members of a same terrorist group, which also determines q, the probability that at least one low-profile group member is found during the probing runs. We consider three possible updating techniques to facilitate our discussion.

First, we define some intermediate probabilities. Let $C_{i|i-1}$ be the probability that the i^{th} probe is selected by the profiling system given that the previous $i-1$ probes were all selected for additional screening, either by the profiling system or at random (with $C_{1|0} = C$). Related to this value is the value $P_{i|i-1}$, the probability that the i^{th} probe is selected *either* by the profiling system *or* at random, given that the previous $i-1$ probes were all selected for additional screening, either by the profiling system or at random. Clearly, $P_{i|i-1} = C_{i|i-1} + (1 - C_{i|i-1})r$. Lastly, let P_i be the total probability that the first i probes are all selected for additional screening either by the PSS or at random: $P_i = \prod_{k=1}^{i} P_{k|k-1}$. Because q is the probability that at least one low-profile member is found amongst the n, q is therefore the probability that not all n members are selected, or $q = 1 - P_n$.

We now find these values under three different updating schemes.

- **Independence:** In this scheme, each member has the same probability of being selected by the PSS, independently of the other members. $C^I_{i|i-1} = C$ and $P^I_{i|i-1} =$

$C + (1 - C)r$ (the superscript I denotes the Independence updating scheme). In this case, the probing flights constitute independent Bernoulli trials, similar to the assumption made in [28], and the probability, q^I of finding at least one low-profile member is

$$\begin{aligned} q^I &= 1 - P_n^I \\ &= 1 - (C + (1 - C)r)^n. \end{aligned} \tag{3.6}$$

While different terrorists are assumed independent by this method, if no low-profile member is found during the probings, then in an attack, the attacker will have been previously flagged during the probing flights. Because he could have been flagged either by the PSS or at random, the conditional probability, $\hat{C^I}$, that he was selected by the PSS (and thus would again be selected by the PSS during the attack) is given by Bayes' Rule:

$$\begin{aligned} \hat{C^I} &= P(\text{selected by PSS}|\text{selected}) \\ &= \frac{P(\text{selected}|\text{selected by PSS})P(\text{selected by PSS})}{P(\text{selected})} \\ &= \frac{1 * C}{C + (1 - C)r}. \end{aligned} \tag{3.7}$$

Under an assumption of independence, q^I and $\hat{C^I}$ are the values of q and $\hat{C}$ to use in Equations (3.3)-(3.5). We note that as n gets large, the probability, q^I, of finding a low-profile member approaches one as we would expect from Bernoulli trials.

- **Maximum Dependence:** On the opposite end of the spectrum from the Independence scheme, one could assume that the system is perfectly able to detect ties between terrorists, such that one terrorist being selected by the profiling system implies all others will be, and one terrorist not being selected implies no others will be. In essence, we treat the n members as if they were a single person conducting n trials.

  Because all members are assumed identical in this case, the i^{th} member will be flagged by the profiling system only if the $i-1^{st}$ person was. So $C_{i|i-1}^{MD}$, the probability that the i^{th} is selected by the PSS given that the first $i-1$ members received secondary screening, under a Maximum Dependence scheme, is the conditional probability that the $i - 1^{st}$ probe was selected by the PSS and not at random, or $C_{i|i-1}^{MD} = \frac{C_{i-1|i-2}^{MD}}{C_{i-1|i-2}^{MD} + (1 - C_{i-1|i-2}^{MD})r}$. Starting with $C_{1|0}^{MD} = C$, we find by induction that $C_{i|i-1}^{MD} = \frac{C}{C + (1-C)r^{i-1}}$. It is essentially the conditional probability of being considered high-risk, given that one was selected for additional screening $i - 1$ times. The total probability that the i^{th} person will be selected for secondary screening given that the previous $i - 1$ members were is therefore $P_{i|i-1}^{MD} = \frac{C + (1-C)r^i}{C + (1-C)r^{i-1}}$, and $P_i^{MD} = \prod_{k=1}^{i} P_{k|k-1}^{MD} = C + (1 - C)r^i$. Because $\hat{C}$ is the probability of the chosen attacker being flagged by the PSS after all n members received secondary screening during the probing flights, $\hat{C^{MD}}$ is equivalent to $C_{n+1|n}^{MD}$

above:

$$\hat{C^{MD}} = \frac{C}{C + (1-C)r^n}, \tag{3.8}$$

and by the relation given earlier,

$$q^{MD} = 1 - P_n^{MD} = (1-C)(1-r^n). \tag{3.9}$$

This is an intuitive result. Because this Maximum Dependence case assumes the terrorists have identical risk levels in the profiling system, if one terrorist is considered high-risk, they all are, and if one terrorist is considered low-risk, they all are. So the only way for a guaranteed low-profile member to be found is if all n members have a low profiling score, occurring with probability $1-C$, *and* if they are not all selected for random screening, occurring with probability $1-r^n$. In contrast to the Independence case, as n gets large, the probability, q^{MD}, of finding a low-profile member does *not* approach 1, but rather the value $1-C$. The likelihood, $\hat{C^{MD}}$, of the attacker being selected by the profiling system given that all n members were flagged during the probing runs approaches 1. The more group members that are selected for secondary screening during the probing runs, the more likely it is that the attacker will also be selected.

- **Average:** A profiling system is unlikely to completely ignore ties between terrorists such that they are completely independent, nor to perfectly detect ties between them such that the selection of one implies the selection of all others. To consider a case somewhere between the Independence and Maximum Dependence cases, we take the straight Average of the two as a third case. At the i^{th} probing run, if the previous $i-1$ members were all selected for secondary screening, we let $\hat{C_i^A}$ be equal to the average of $\hat{C_i^I}$ and $\hat{C_i^{MD}}$:

$$\hat{C_i^A} = \frac{C_i^I + C_i^{MD}}{2}. \tag{3.10}$$

Similarly we let conditional probability of the i^{th} member being flagged by the PSS given that the first $i-1$ probes received secondary screening to be the average of those under the Independent and Maximum Dependence cases: $C_{i|i-1}^A = \frac{C_{i|i-1}^I + C_{i|i-1}^{MD}}{2}$, which also causes $P_{i|i-1}^A = \frac{P_{i|i-1}^I + P_{i|i-1}^{MD}}{2}$. Using the definition $P_n^A = \prod_{i=1}^n P_{i|i-1}^A$, we can find $q^A = 1 - P_n^A$ iteratively.

We will use the Averaging technique in the remainder of the analysis, except if indicated otherwise.

3.4 Analysis techniques

To evaluate the probability that the terrorists attempt and succeed at their attack, we need to estimate the model's parameters. However for the most part, the values are either not publicly available or depend on the particular plot being considered by the terrorists. We will, therefore, discuss some possible estimates, based on publicly available information, and evaluate the model considering a wider range of values.

The effectiveness of the primary and secondary screening, p_1 and p_2, depends heavily on the exact nature of the terrorist plot, as discussed earlier. Furthermore, performance specifications for metal detectors and x-ray bag scanners in use in the U.S. are kept classified. However, reports from the U.S. Government Accountability Office (formerly the General Accounting Office) indicate that screeners are roughly 80% likely to detect certain unauthorized items, although it is unclear whether this at the primary screening level, secondary screening level, or both. This assessment is based on a 1987 inspection of screeners that revealed that 20% of dangerous objects were not detected. Subsequent inspections in 1999, 2003 and 2005 indicated that performance was not improving and may even be worsening [90, 138, 146, 148]. Furthermore, while these results may correspond to the detection of guns, knives or even explosives contained in carry-on luggage, currently neither primary nor secondary screening is likely to detect explosives carried on a person's body, although a new walk-through detector known as a "sniffer", which blows a puff of air on the passenger to dislodge trace particles of explosives that it can then detect, might increase the value of p_1 (and hence p_2) for such plots.

The estimated ability of the PSS to recognize a high-risk individual, as measured by C, is also not publicly available. Moreover, its value could depend on whether the group members involved in the plot are well-known terrorists or new recruits. In [157], the authors define a CAPPS multiplier, β, to be the ratio of the proportion of threats amongst selectees versus non-selectees. If T is the event that a bag contains a threat, R the event that the passenger is a high-risk individual (a selectee) and $\overline{R}$ the complement event, then

$$\beta = \frac{P(T|R)}{P(T|\overline{R})}.$$

β gives an indication of how much more likely those selected by the PSS are to have a dangerous object than non-selectees. If the PSS is working correctly, β should be greater than 1. We can rearrange the terms above to get

$$\beta = \frac{P(R|T)P(\overline{R})}{P(\overline{R}|T)P(R)}. \tag{3.11}$$

$P(R|T)$ is the probability of a person being selected for additional screening given that he is carrying a threat object, or C in our model. $P(R)$ is the total fraction of passengers considered to be selectees, which is currently estimated to be 15%, but changes to Secure Flight might decrease it to 5%. The study in [157] was conducted with the Federal Aviation

Administration, and while the authors had access to classified parameter values such as β, they altered the values for their published paper and assigned the value $\beta = 155$ to use in their analysis. If we use this value of β and solve for C in Equation (3.11), we find an estimate for C of 96% under the current PSS where 15% of passengers are selected, and 90% under the proposed Secure Flight where fewer would be selected. However, on September 11, only six out of the nineteen terrorists were flagged by CAPPS [47], suggesting a value for C of only 32%. As we will see later, uncertainty in the parameter C can yield conflicting conclusions.

r, the fraction of lower-risk passengers selected at random for additional screening, is believed to range between 5-10%. Lastly, the parameter τ is the most difficult to estimate because it measures a particular terrorist cell's willingness to attempt an attack given the likelihood and possible consequences of getting caught. We have little understanding of how to estimate deterrence levels so we will refrain from doing so, choosing instead to explore the probability of a successful attempt over a wide range of values of τ.

Given the uncertainty of the above estimates, we follow the example of [35] and consider a wide range of reasonable values for each parameter in our analysis. We can then examine how the probability of a successful terrorist attempt varies with the parameters, and determine which parameters are most influential in preventing a successful attack.

- C ranges between 0 (the terrorist will never be selected by the PSS) and 1 (the PSS will always select him), in units of 0.1.
- r ranges between 0 and 0.20, in units of 0.05. Selecting more than 20% of passengers at random would not be realistic given the goal of reducing the amount of screening for presumed innocent passengers.
- p_1, the effectiveness of the primary screening, will range from 0% to 100% effective, in 10% increments.
- p_2, the effectiveness of the secondary screening, ranges from p_1 to 100%, in 10% increments.
- τ will be allowed to range between 0, in the case where the terrorists would attempt the plot no matter what, and 1, in the case where the terrorists would attempt the plot only if guaranteed success, in 0.05 increments.

For each combination of these values, which we call a *scenario*, we calculate the probability of a successful attempt, according to Equation (3.5), using the Averaging method for updating C, assuming a group of five members. (In Section 3.5.2, we will also discuss how the choice of updating method and assumptions on the number of terrorists influence the results).

Because we don't know which scenarios are more likely to occur than the others, we weight each one equally when aggregating similar scenarios to create an average value for $P(AS)$. This assumption is not without its flaws, however. The restriction $p_2 \geq p_1$ causes certain types of scenarios to carry greater weight than others in our analysis. For instance,

when $p_1 = 0$, p_2 can take on any value between 0 and 1, but when $p_1 = 1$, p_2 can equal only 1, corresponding to far fewer scenarios. Thus when we take averages, there is more variance amongst scenarios in which p_1=0 than those in which p_1=1. There is a symmetric problem if we fix p_2 and examine the corresponding valid range of p_1. It seems reasonable that certain scenarios should be more likely to occur than others. However, since this probability distribution is unknown to us, we use a uniform prior probability distribution and average equally over all scenarios. Individual scenarios will be considered when greater subtlety is warranted.

3.5 Interpreting the model

We examine now the influence each parameter has on the likelihood of a successful terrorist attack and relate these conclusions to the original hypotheses of supporters and critics of passenger pre-screening systems.

3.5.1 Effects of screening effectiveness parameters on attack success

The arguments raised by supporters of profiling systems rely heavily on the assumption that the profiling system will be very good at identifying terrorists (i.e. that $C \approx 1$). Admiral Loy claimed that "CAPPS II will dramatically enhance customer service by identifying the vast majority of air travelers as innocent passengers" [143]. But they do not say precisely why they are confident. Not all potentially dangerous individuals appear on terrorist watch lists, and data-mining algorithms meant to sort out suspicious behavior need not be especially successful. [20] notes that at the time of the 2002 Washington sniper crisis, historical data mining about serial killers led to the widespread belief that the perpetrators were white, when in fact, they were not.

However, even if we temporarily assume the profiling system to be accurate, placing faith in passenger screening solely based on the performance of the profiling system may constitute wishful and incomplete thinking. Aggregating scenarios into (p_1, p_2) groupings of twenty percentage points each, we found the *average* probability of a successful attack attempt, $P(AS)$, over all scenarios within each group. Figure 3-4 shows the relationship between this average probability of a successful attack and the effectiveness of the primary and secondary levels of screening, when $n = 5$ and $n = 10$ and for each of the updating schemes discussed earlier. As expected, for all six of the charts, $P(AS)$ decreased drastically in scenarios where screening was nearly perfect (p_1 and p_2 both very high), but was perhaps uncomfortably high for even moderate values of p_1 and p_2. For instance, in most of the charts, if the primary screening was less than 60% effective, there was at least a 10% chance on average that the plot would be attempted and successful, regardless of the effectiveness of secondary screening.

This varies, however, by the number of terrorists in a group and the ability of the profiling system to detect ties between terrorists. Comparing the two columns of charts for $n = 5$ and

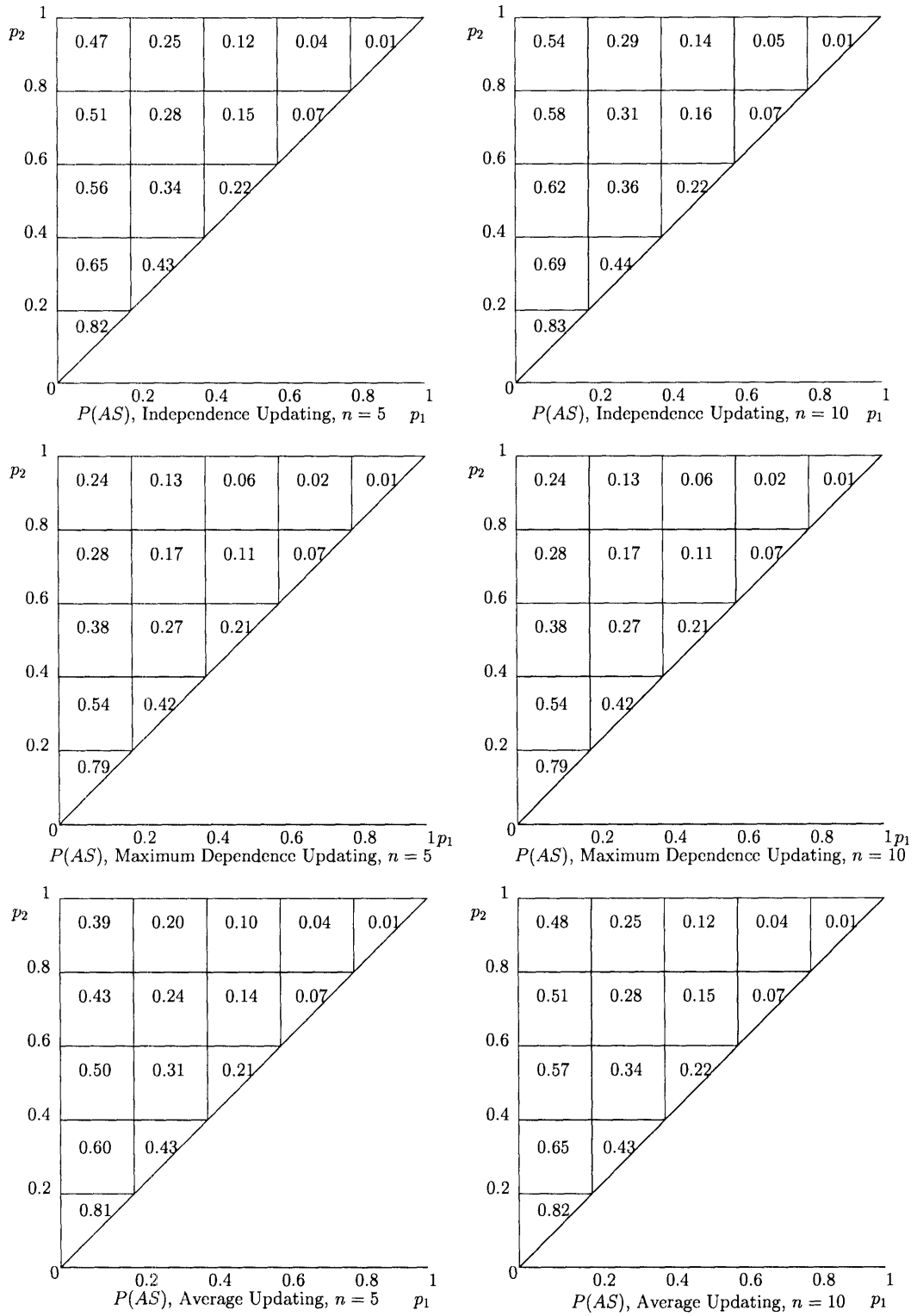

Figure 3-4: The probability of a successful attack attempt, $P(AS)$, by (p_1, p_2) range, averaged over all other parameter values, under $n = 5$ and $n = 10$ and the three updating schemes used: Independence, Maximum Dependence and Average updating.

$n = 10$, we see that having a larger terrorist group generally increased the probability of a successful attempt in the Independence and Average updating schemes but did not change the chance of success within two significant digits for the case of Maximum Dependence. It can be shown that as n increases, the probability of a successful attack is non-increasing under Maximum Dependence updating, as will be discussed in Section 3.5.2. When the terrorists are treated independently by the profiling system, then having a larger pool from which to select an attacker increases the likelihood that at least one low-profile member will be identified during the probing trials. However, when the profiling system is very good at detecting links between terrorists such that one member's high-risk rating implies a high-risk rating for the other members, then having more members does not offer any benefit. The Average updating scheme is somewhere in between the two, and therefore having additional members in the group helps somewhat. The amount by which the success probability increases from $n = 5$ to $n = 10$ under the Independence updating scheme also depends on the effectiveness of the primary and secondary screening. When there is little difference between primary and secondary screening (either both are poor or both are very good) then there is less of an advantage for the terrorists in avoiding secondary screening. Having additional members in the group does not increase the average value of $P(AS)$ significantly. By contrast, when secondary screening is much better than primary screening (as in the upper left-hand corner of the charts), then avoiding the secondary screening by sending additional probes to find a low-profile member is more beneficial to the terrorists.

If we compare the average probability of a successful attempt between updating schemes, we see that a profiling system capable of detecting ties between terrorists, as in the Maximum Dependence scheme, yields lower probabilities of attack than a system that selects terrorists independently. This is particularly true in the region of the chart corresponding to low primary screening effectiveness and high secondary screening effectiveness. Once again, having a profiling system that is more robust to probing (as would be the case if the profiling system can detect ties between terrorists) reduces $P(AS)$ the most when the secondary screening that selectees receive is significantly better than the primary level of screening.

However, is this robustness possible within the framework of civil liberties? The original CAPPS algorithm supposedly bases its decision on passenger record information collected by the airlines, such as class of ticket, method of payment, or whether the flight is one-way or round-trip. While a group of terrorists coordinating an attack might share these characteristics, these are also simple characteristics that apply to many low-risk passengers as well. A higher-performing profiling system, capable of identifying ties between terrorists, would likely need access to more personal information, which raises questions of whether the collection of such information is a violation of civil liberties. As pointed out in [20], after many revisions to the algorithm in response to criticism, the profiling system that remains might not be as effective at detecting ties between terrorists as we might hope. The system would then be more vulnerable to probing and other loopholes, limiting its value.

Even if the profiling system is quite effective, we still might not be as safe as we think, as the next example shows. Suppose that the PSS is 100% effective at identifying terrorists ($C = 1$). As seen in Equation (3.1), the terrorists' probability of success in this case,

equal to $1 - p_2$, depends entirely on the effectiveness of the secondary screening. If their plot involves the use of weapons not currently prohibited from aircraft or not detectable by the secondary screening process, then their success is guaranteed, *despite a perfect profiling system*! Supporters of profiling never speak explicitly of the effectiveness of screening imposed on selectees. One gets the impression they tacitly assume that secondary screening is highly effective (i.e. that $p_2 \approx 1$). But evidence for that viewpoint is limited. On September 11 itself, six of the nineteen terrorists were subjected to additional scrutiny, but not one was stopped from boarding the aircraft. At the time, the additional screening consisted of *checked luggage* screening and did not search for (let alone forbid) the box-cutters that were apparently instrumental to the plot.

Secondary screening today, of course, is more demanding, but its effectiveness has been sharply questioned. A recent report of the (then) Inspector General to the Department of Homeland Security, Clark Kent Ervin, was not encouraging. Undercover tests conducted in 2003 revealed weaknesses in employee training; screening equipment and technology; security policy and procedures; and management and supervision [136]. According to the report, the passenger screening process in place at fifteen airports repeatedly failed to detect weapons and explosive materials. Even worse, a report released in 2005 indicated that the performance still had not improved and likely would not improve without advanced technology [90, 138]. According to DHS Inspector General Richard Skinner, the screeners "fared no better than the performance of screeners prior to September 11, 2001" [46]. And in February 2005 the country's confidence was further shaken when, in two separate incidents at New Jersey's Newark International Airport, a butcher knife and a fake test bomb made their ways onto the plane (the fake bomb went even as far as Amsterdam) [99]. Representative John Mica, chairman of the subcommittee, declared the results of the tests to be "bad enough" for general screening and "absolutely horrendous" with respect to detecting explosives [58]. The main screening devices used at checkpoints are an x-ray machine and a metal detector, but as Representative Peter DeFazio of the House Aviation Subcommittee noted, "You're not going to find plastic explosives with a metal detector, no matter how hard you try" [58]. For certain types of plots, the true value of p_2 might be far below one. In that circumstance, directing the terrorists to secondary screening might be a hollow victory.

The optimists also seem indifferent to p_1, the detection rate for primary screening. But if $C < 1$, we cannot ignore this parameter. As discussed earlier, terrorists could use innocent passengers in an attack without their knowledge or find a low-risk terrorist through probing. Thus, if C is low, any inadequacies of primary screening may substantially raise the chance of a successful attack. Figure 3-5 shows the smallest values of C (amongst those explored), by (p_1, p_2) grouping, in which the average probability of a successful attempt would be less than 20%, 10% or 5%. An entry of "None" indicates that no value of C yielded an average value of $P(AS)$ sufficiently low over scenarios in that group. If we interpret 20%, 10% and 5% as tolerance levels for acceptable risks of attack, we see first that as the acceptable risk of attack decreases from 20% to 5%, we require substantial improvements in the quality of secondary screening and in the quality of the profiling system to reduce the chance of attack below those levels. We see that when the primary screening is fairly effective, then our

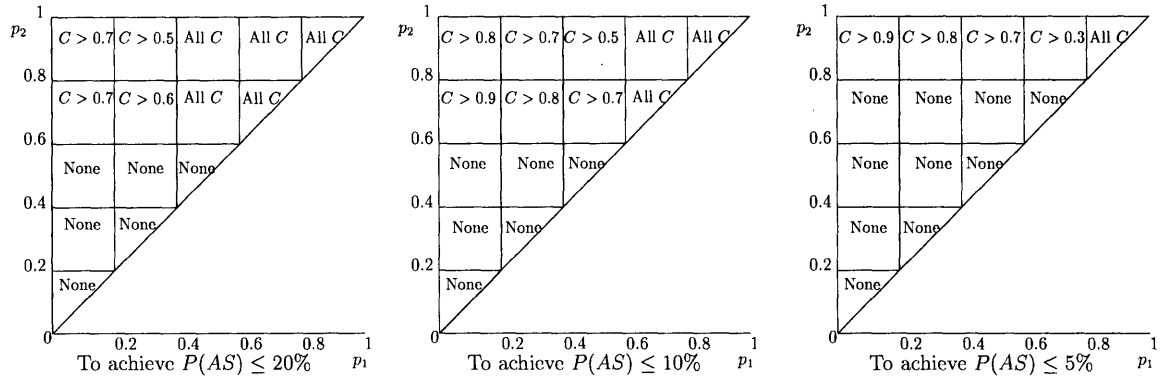

Figure 3-5: The minimum profiling effectiveness (C) required to experience an average $P(AS)$ less than or equal to 20%, 10% or 5%, by (p_1, p_2) grouping.

dependence on an effective profiling system is minimal. Over the set of scenarios in which the primary screening had an 80-100% chance of detecting the plot (corresponding to the upper right-hand quadrant of the charts), a successful attack would occur with less than 5% probability (when averaged over r and τ), regardless of the profiling system's effectiveness. Scenarios in which p_1 ranged between 60% and 80% yielded an average $P(AS)$ between 5% and 10% regardless of the value of C, and an average $P(AS)$ less than 5% if the secondary screening was in the 80-100% effectiveness grouping and the profiling system at least 40% effective. Thus, stringent primary screening might be costly and time-consuming but could render profiling unnecessary.

For the groupings in which both the primary and secondary screening were less than 60% effective (lower left-hand quadrant), the probability of a plot being successfully attempted was found to be greater than 20%, even when considering scenarios in which the profiling system perfectly selected all terrorists. In short, having ineffective primary screening could be dangerous even if the PSS is effective at identifying terrorists and especially if it is not. Even if we assume highly effective secondary screening, we may need a very effective profiling system to be comfortable with a low level of primary screening.

Figure 3-5 shows the same trend seen in Figure 3-4 that if both primary and secondary screening have similar detection rates, then profiling offers little value. When both p_1 and p_2 are very high, the average probability of a successful attack is fairly low, regardless of the quality of the profiling system. Similarly, when both levels of screening are ineffective, then not even a perfect profiling system can substantially reduce the threat of attack. It is only when the secondary screening is significantly better than the primary screening that profiling appreciably decreases the terrorists' success rate.

Thus, though many supporters believe the pre-screening systems will be effective at identifying terrorists, this alone is not sufficient reason to be optimistic about their overall ability to prevent terrorism. Without explicitly considering the values of p_1, p_2, or $p_2 - p_1$, they miss the point that identifying high-risk people is beneficial only if that capability

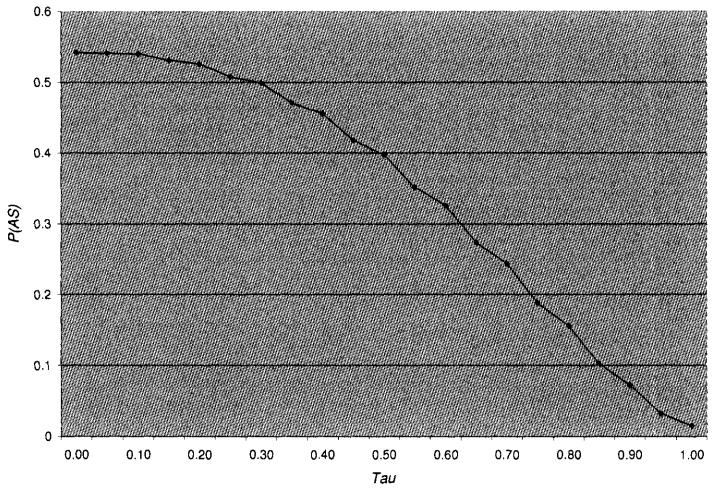

Figure 3-6: The probability of a successful terrorist attempt as the deterrence threshold τ varies from 0 (never deterred) to 1 (deterred unless guaranteed success), averaged over all scenarios under $n = 5$ and Averaged updating.

reduces the chance that an attack would succeed.

3.5.2 Deterrence effects

While C, p_1, and p_2 may indeed be far below one, are they still high enough to convince the terrorists to stay away? Supporters of profiling systems rely at least partially on deterrence effects for their optimism, while critics such as Chakrabarti and Strauss ignore the possibility that terrorists might be deterred. How important is the deterrence threshold, τ?

Figure 3-6 shows the relationship between the average likelihood, $P(AS)$, of a successful terrorist attack and the deterrence threshold, τ. The average $P(AS)$ decreases as the terrorists require a higher chance of success in order to attack, as we would expect, and is also concave in τ. As τ initially increases from zero, those plots first to be abandoned are those that were unlikely to succeed in the first place. Because they contribute little to the total probability of success, their abandonment as τ initially increases causes only a slight decrease in the probability of success. However, as τ approaches 1, the average probability of success falls close to zero.

Given its influence on the average likelihood of a successful attempt, we would like to explore how deterrence affects probing. The primary argument used in [28] is that by probing, the terrorists are guaranteed to find somebody who will slip through the system, thus assuring a probability of success equal to S, rather than $\hat{S}$. First, they assume an arbitrarily large group of independent terrorists. But are the groups really arbitrarily large? Next, the assumption that group members are each classified independently as high- or low-risk

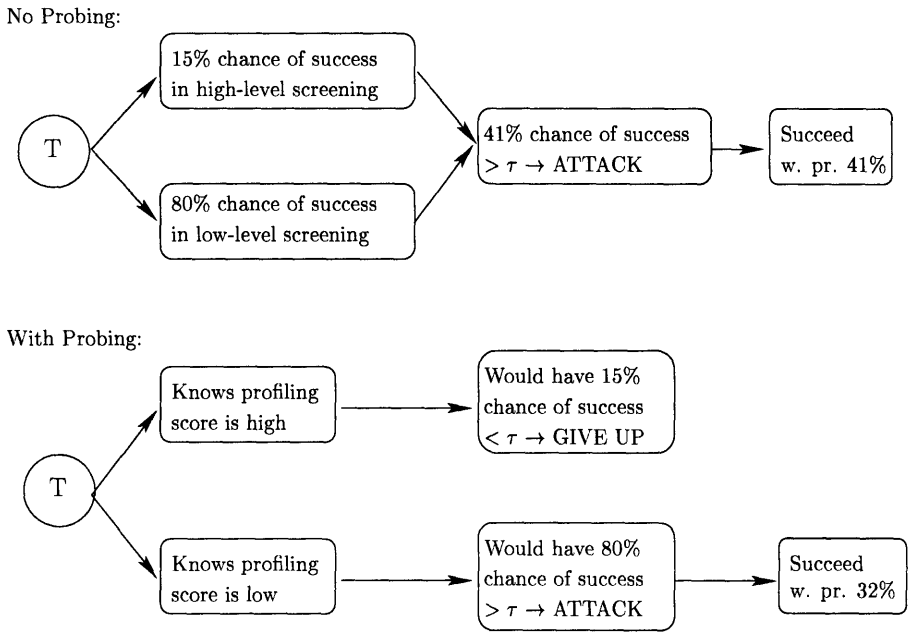

Figure 3-7: Probing the system can sometimes discourage rather than reassure the terrorist. ($p_1 = 0.2, p_2 = 0.85, r = 0, C = 0.6, \tau = 0.25$)

suggests that U.S. intelligence services are unable to recognize connections between dangerous individuals. This could be true, but models that effectively assume independence in probabilistic calculations could substantially overestimate the chance of finding a group member who gets only primary screening. Perhaps most important, however, is the point that terrorists might cancel their plans if they get evidence of an unacceptable probability of success. Probing might provide that evidence.

Consider a lone terrorist, and suppose that he has made the following (presumably accurate) estimates about the system:

- Primary screening is 20% effective at detecting his plot, ($p_1 = 0.20$),
- Secondary screening is 85% effective at detecting his plot ($p_2 = 0.85$),
- No passengers are selected at random for additional screening ($r = 0$),
- The PSS has a 60% chance of selecting the terrorist for secondary screening ($C = 0.60$).

Assume that the terrorist will not attack unless he has at least a 25% chance of success ($\tau = 0.25$). In Figure 3-7, we compare the terrorists chance of success if he does not probe the system before attacking to that if he does probe. If he does not probe, he first compares his estimated probability of success with τ: With 60% probability, he would undergo secondary screening at the time of the attack and thus would have a 15% chance of success. Otherwise,

he would pass through primary screening and experience an 80% chance of success. His total chance of success is then (0.60)(0.15)+(0.40)(0.80) = 41%, which is larger than τ. He decides to attack, and as per his calculations, succeeds with probability 41%.

However, if he does probe, then he waits to decide whether or not to attack until after the probing run. During this trial there are two possibilities:

1. With 60% probability, he is selected for secondary screening, knows he is classified as high-risk and updates his estimate of C to 1 (This is just the conditional probability that he was flagged by the PSS and not at random, as we have assumed there is no random screening in this example). Knowing that if he were to attack, he would face secondary screening and succeed with only 15% probability, which is less than τ, he gives up without attacking.

2. With 40% probability, he avoids secondary screening during his trial run and knows for certain that he is considered low-risk (he updates C to 0). If he were to attack, his probability of success under primary screening would be 80%. Because this is higher than his deterrence level τ, he decides to attack.

Because he attempts the attack only if he is deemed low-risk during the probing run, which happens with 40% probability, and subsequently succeeds with 80% probability, his overall chance of success if he probes first is only 32%, which is lower than if he had not probed.

What causes this? As the September 11 Commission [107] explained, "Terrorists should perceive that potential targets are defended. They may be deterred by a significant chance of failure". When terrorists probe to gain information about the system, the information they get might be discouraging, and they might cancel an attack that otherwise would have had an appreciable chance of success. Unless one believes that terrorists cannot be deterred (the widely accepted opinion is otherwise. See, e.g., [41].), analyses that treat deterrence as a negligible phenomenon might be too pessimistic. We will now show mathematically conditions under which probing does *not* improve the terrorists' chance of success.

Probing decreases chance of success, n = 1

For the case of a lone terrorist, we can show that the probability of success is *never* higher when the terrorist probes than when he does not probe. Let $P(AS)_{probe}$ be the probability of a successful attempt when the lone terrorist first probes the system, and $P(AS)_{noprobe}$ the same probability when he attacks directly without first probing. Let

$$A = (1-C)(1-r)(1-p_1) + (C+(1-C)r)(1-p_2), \tag{3.12}$$

then $P(AS)_{noprobe}$ equals A when $A \geq \tau$ and 0 otherwise. $P(AS)_{probe}$ is as in Equation (3.5), with $q = (1-C)(1-r)$, S and $\hat{S}$ are as in Equations (3.2) and (3.3), and $\hat{C} = \frac{C}{C+(1-C)r}$ (Because there is only one terrorist, the question of selecting an updating scheme for finding q and $\hat{C}$ is moot).

Using the fact that $p_2 \geq p_1$, then $S \geq A \geq \hat{S}$ and $A \geq qS$, giving us four cases to consider:

- **Case 1:** $S \geq A \geq \hat{S} \geq \tau$

$$\begin{aligned}
& P(AS)_{probe} - P(AS)_{noprobe} \\
= \; & (qS + (1-q)\hat{S}) - A \\
= \; & (1-C)(1-r)\left[(1-r)(1-p_1) + r(1-p_2)\right] \\
+ \; & (C + (1-C)r)\left[\frac{(1-C)r}{C+(1-C)r}(1-r)(1-p_1) + \frac{C+(1-C)r^2}{C+(1-C)r}(1-p_2)\right] \\
- \; & \left[(1-r)(1-C)(1-p_1) + (C+(1-C)r)(1-p_2)\right] \\
= \; & 0,
\end{aligned}$$

- **Case 2:** $A \geq qS \geq \tau, \hat{S} < \tau$

$$\begin{aligned}
& P(AS)_{probe} - P(AS)_{noprobe} \\
= \; & qS - A \\
\leq \; & 0,
\end{aligned}$$

- **Case 3:** $A \geq \tau, qS < \tau, \hat{S} < \tau$

$$\begin{aligned}
& P(AS)_{probe} - P(AS)_{noprobe} \\
= \; & 0 - A \\
\leq \; & 0,
\end{aligned}$$

- **Case 4:** $A < \tau, qS < \tau, \hat{S} < \tau$

$$\begin{aligned}
& P(AS)_{probe} - P(AS)_{noprobe} \\
= \; & 0 - 0 \\
\leq \; & 0.
\end{aligned}$$

In each of the four cases, the probability of a successful attempt when the terrorist first probes the system is always less than or equal to that if he does not probe the system. If he attacks directly without first probing then he might succeed, even in circumstances that would have caused him to give up had he first probed.

Probing decreases chance of success, n > 1, Maximum Dependence

Similar statements can be made in certain cases involving multiple terrorists. For instance, if the Maximum Dependence updating scheme of Section 3.3.3 is assumed, then regardless of the number of terrorists, the probability of a successful attempt is lower if they probe the system first than if they select a member arbitrarily amongst the group of n and send this member on the attack directly. The proof of this is similar to that above for $n = 1$, but we use $q = (1-C)(1-r^n)$ and $\hat{C} = \frac{C}{C+(1-C)r^n}$, as described in Section 3.3.3.

A related observation is that the probability of a successful attack under the Maximum Dependence assumption actually *decreases* as the size, n, of the terrorist group *increases*. Recall that the Maximum Dependence scheme is equivalent to sending a same person on n consecutive probing flights. If this person is considered high-risk, then on each of these flights he will be selected for additional screening. Had he probed only once, then perhaps he might believe he was pulled aside at random rather than due to a high-risk rating. However, after n tries, particularly as n gets large, he becomes more convinced of his high-risk rating and more pessimistic about his chances of success. Thus, as n gets larger, he is more likely to give up on the attack. This is the opposite effect than that seen when terrorists are assumed independent.

Probing may increase or decrease chance of success, n > 1, Independent terrorists

If we have reason to believe that terrorists are selected independently of one another by the profiling system, then there exist examples both where probing helps and where it does not help the terrorist. While we saw in Section 3.3.3 that the likelihood of finding a low-profile member approaches 1 as the size, n, of the terrorist group gets very large, for moderate values of n, the group might become too discouraged to attempt the attack.

To see this, we consider two examples. Once again we have $P(AS)_{noprobe}$ equal to A when $A \geq \tau$ and 0 otherwise, where A is as in Equation (3.12). $P(AS)_{probe}$ is again as in Equation (3.5), but now with $q = 1 - [C + (1-C)r]^n$ and $\hat{C} = \frac{C}{C+(1-C)r^n}$, as in Equations (3.6) and (3.7).

- **Example 1** $p_1 = 0.2, p_2 = 0.5, C = 0.8, r = 0.01, \tau = 0.55, n = 5$

  If a terrorist were to commit the attack without probing, his probability of success would be $A = 0.5594$. Because this is greater than τ, he commits the attack, and $P(AS)_{noprobe} = 0.5594$. If, instead, the group considers probing, their probability of success if they find a low-profile member, S, would be 0.797 and greater than τ, but their probability of success if they must use a previously-flagged member, $\hat{S}$, would be 0.5007 and less than τ. If they decide to probe, they will commit the attack only if they find a low-profile member. However, their probability q of doing so is only 0.6682, and their probability of success is therefore $qS = 0.5326$. Because this is less than τ, they actually decide to give up without probing, and thus $P(AS)_{probe} = 0$, which is less than had they attacked directly, without considering probing.

- **Example 2** $p_1 = 0.2, p_2 = 0.5, C = 0.3, r = 0.01, \tau = 0.55, n = 5$

  In this example, if a terrorist were to commit the attack without probing, his probability of success would be $A = 0.7079$. Again, this is greater than τ, so he will commit the attack, and $P(AS)_{noprobe} = 0.7079$. If the group considers probing, their probability of success using a low-profile member, S, would be again 0.797 and greater than τ, and their probability of success using a previously-flagged member, $\hat{S}$, would be 0.5068 and still less than τ. Again, they will commit the attack only if they find a low-profile member, but this now occurs with a greater probability, q, since C is smaller

in this example. We have $qS = 0.7948 \geq \tau$, so they decide to probe and will attempt and succeed with their plot with probability $P(AS)_{probe} = 0.7948$. Here we have $P(AS)_{probe} > P(AS)_{noprobe}$.

In the first example, the likelihood of finding a low-profile member, q, and the chance of success using a high-profile member were too low to justify probing, causing the terrorists to give up. The terrorists' chance of success would be higher proceeding directly with the attack than if they probed first. In the second example, the probability of being selected by the PSS was lower, improving the terrorists' chances of finding a low-profile member during the probing runs. In this case, conducting probing flights would improve their chance of success.

This highlights a few important conclusions. First, excluding deterrence from an evaluation of profiling systems can create an overly pessimistic assessment of such systems. Law enforcement measures (whether against terrorism or other crimes) serve not only the purpose of intervention during an infraction but also the purpose of deterring such infractions. Such capability must be considered and modeled.

A second conclusion is that different assumptions about how the profiling system selects terrorists can influence the system's vulnerability to potential loopholes. If the system is unable to detect ties between terrorists, such that they are selected independently of one another for secondary screening, then the probing behavior outlined by Chakrabarti and Strauss can, at least in some cases, cause the terrorists to achieve a higher chance of success, particularly if they have a large pool of members from which to choose an attacker. However, as the system becomes better able to detect relationships between members of a same group, then the probing flights might discourage the terrorists if they realize the system is more hardened than they initially thought.

3.5.3 Role of random screening

Lastly, we examine the role of random screening. Many supporters of random screening argue that it acts as a deterrent: by keeping the passenger screening process even moderately unpredictable, terrorists will be less able to game the system to their benefit and might get discouraged from attempting an attack. Figure 3-8 shows the relationship between the average $P(AS)$ over all other parameters and r. We see that the curve is roughly linear and decreasing in r, but with a shallow slope: when $r = 0$, the average $P(AS)$ is near 0.38, and when $r = 20\%$, this decreases by less than seven percentage points to around 0.31. This suggests that random screening, at the relatively small level that airports could conduct it without frustrating passengers, is not an influential parameter in reducing the likelihood of a successful attack. Because our system does not want to screen too many "innocent" passengers, we are forced to keep r low, and as such, its effect is limited.

Furthermore, terrorists will not fear random screening if the additional screening is not able to detect their plot. Similar to the PSS effectiveness parameter C, random screening yielded a greater effect at reducing the risk of a successful attack in scenarios where p_2 was high relative to p_1 than when the two screening levels had similar effectiveness.

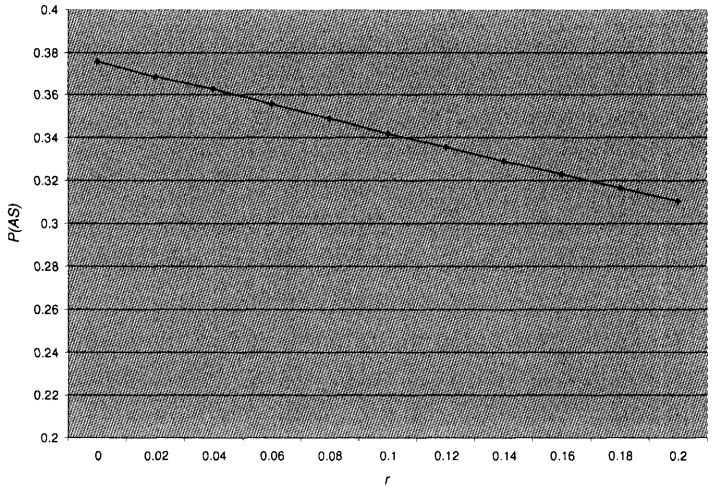

Figure 3-8: The probability, $P(AS)$ of a successful terrorist attempt as a function of the percentage, r, of randomly selected low-profile passengers, averaged over all possible parameter values.

3.6 Conclusions and policy implications

There have been conflicting interpretations of computerized pre-screening systems. Optimists feel that profiling is the "right answer" and will facilitate the screening process for the majority of passengers believed to be low-risk. Pessimists feel that using a profiling system in the case of personal pre-boarding screening will allow terrorists to obtain information about their status that they can use to their advantage. Both factions' arguments may be somewhat shortsighted, however.

Critics of profiling systems may have given insufficient weight to deterrence (τ), because of which the selection and screening system, although imperfect themselves, might still prevent attacks from being attempted. Moreover, although we assumed that the terrorists' estimates of system parameters match the parameter's true values, this does not have to hold in reality. If terrorists believe the values of C, p_1 or p_2 are higher than they actually are, then this belief alone might discourage them. We have also seen that probing the system could sometimes prevent a terrorist act rather than ensure its success.

Supporters of such systems have focused mostly on the ability of the algorithm to identify terrorists (C) and say little about screening effectiveness of both lower-risk passengers and selectees (p_1 and p_2). Depending on the true value of C, our conclusions on where to focus security efforts may change. For instance, the work of [157] used an estimate of C equivalent to roughly 90-96%, while the percentage achieved on September 11 was only 32%. Figure 3-5 shows that if $C = 0.32$, then primary screening must be more than 40% effective and secondary screening more than 60% effective in order to attain an average probability of attack less than 20%, while if $C = 0.96$, only the secondary screening is relevant. To have

less than a 5% chance of attack, both primary and secondary screening must be at least 80% effective if $C = 0.32$, and again only the secondary screening is relevant if $C = 0.96$. If we are unsure of the effectiveness of the profiling system, then it is necessary to focus efforts on improving the primary screening that every passenger receives. If we have evidence that the profiling system is effective, then this, too, is insufficient as we must ensure that the secondary screening received by selectees is effective. Having a robust and effective profiling system does not help if the underlying screening is incapable of preventing an attack.

However, once a profiling system is developed, how will we know whether or not it is effective? First, such systems are fundamentally difficult to test. While a security screener operating an x-ray scanner can be tested by sending fake weapons through the x-ray and measuring whether the employee detects them, similar tests are impossible with profiling systems. An algorithm in preliminary testing might achieve a very high likelihood of selecting a terrorist having the characteristics the system has been designed to detect. However in practice, the true characteristics of a terrorist might be more difficult to recognize, and there would be no way of knowing this in advance of an attack. Second, as we have seen in this chapter, a profiling system capable of detecting ties between terrorists is significantly more robust to terrorist loopholes, such as probing, than a system selecting terrorists independently of one another. Yet, limitations on the nature of data that is allowed to be collected might drastically limit the ability of a profiling system to draw such ties, and may further reduce its overall effectiveness. In light of this, in order to increase the likelihood of selecting a terrorist, the system would likely have to cast a wide net and select a greater percentage of passengers in general, thus increasing the "hassle-factor" profiling was expected to reduce.

It appears that profiling systems are neither right answers nor mere carnival games, but instead lie somewhere between the two assessments. Immediately dismissing their benefit as negligible due to the existence of loopholes is short-sighted, since the intricate issues surrounding terrorist behavior and deterrence are difficult to capture in a model, and could set a dangerous precedent for other security measures for which loopholes might exist (one could argue that there exist loopholes around most security measures). Yet, basing the "foundation" of aviation security on a single parameter without consideration for the entire system is likewise myopic. Straightforward mathematical reasoning can help clarify the rhetoric surrounding many of these issues.

Chapter 4

Securing Cargo Holds of Passenger Aircraft: A Cost-Benefit Perspective

In addition to the effectiveness of security systems, another important consideration is their cost. The attacks of September 11 showed us that the costs of a terrorist attack can be greater than previously imagined, and that updating our security systems to protect against a wider range of attacks is likewise expensive. In light of this tradeoff, how do we decide which security systems are worth implementing?

In this chapter, we examine measures to protect cargo compartments on passenger aircraft against Improvised Explosive Devices (IED's), focusing on three avenues by which such a device might find its way onto a passenger aircraft: through checked luggage, via United States Postal Service airmail, or in commercial cargo. Correspondingly, we consider three measures protecting against such attacks. The first, known as Positive Passenger-Bag Match (PPBM), prevents a suitcase from remaining on an aircraft if its accompanying passenger has not boarded the plane by the time of departure. The next is a policy implemented immediately after 9/11 prohibiting United States Postal Service airmail packages weighing more than one pound from passenger aircraft. One concern with this policy, however, is that it might appear contradictory to prohibit airmail packages while significantly larger cargo shipments can be transported on passenger planes largely unscreened. Therefore, we consider a third hypothetical policy, parallel to that for airmail, in which cargo packages would be prohibited from passenger aircraft.

Our objective in this chapter is to explore how to evaluate the cost-effectiveness of a policy in light of the uncertainty surrounding the cost of the policy, cost of attack and threat of attack, and how to draw meaningful conclusions from such analysis. We begin in the following section by discussing the three measures in detail. In Section 4.2, we provide a cost-benefit method for assessing the value of a security policy, and then in Section 4.3 we apply the method in the context of the three policies described above. The results are interpreted in Section 4.4. We find PPBM to be cost-effective even if attacks on checked luggage occur rarely, while the removal of cargo is not cost-effective unless there is reason to believe an attack on cargo is imminent. Although the case of removing airmail is somewhat inconclusive, we are able to demonstrate that the implementation of one security measure

does not set a precedent for instituting another measure protecting against a similar type of attack if the two vary considerably in cost. Thus, it is not necessarily inconsistent to remove airmail packages from planes while continuing to ship cargo. However, there is some potential inaccuracy in focusing on each security measure individually and not considering the system as a whole. In particular, as one possible target is fortified, the threat, rather than disappearing completely, may instead be diverted to a more vulnerable target. Such issues are discussed in Section 4.5. The conclusions are summarized in Section 4.6.

4.1 Three security measures

4.1.1 Positive Passenger-Bag Match

After the bombing of Pan Am Flight 103 over Lockerbie, Scotland, caused by an explosive packed in a checked suitcase unaccompanied by its owner, international flights implemented Positive Passenger-Bag Match (PPBM), a policy that ensures that any bag loaded onto a plane has its accompanying passenger on-board. If the passenger does not board the plane in time, the baggage is "pulled" from the aircraft prior to departure.

Shortly following the September 11, 2001 attacks, Congress mandated in the Aviation and Transportation Security Act (ATSA) [133] that *all* checked luggage on domestic flights be screened for explosives prior to departure from the originating airport. Such screening could take various forms: CT-scanning Explosive Detection Systems (EDS), Explosive Trace Detectors (ETD), bomb-sniffing dogs, hand searches and PPBM. Up to this point, most airports were not screening checked luggage for explosives except in the case of passengers selected by the Computer-Assisted Passenger Pre-Screening System, CAPPS (see Chapter 3). Thus, when the requirement for 100% screening took effect, most airports struggled to acquire the equipment and staffing necessary to attain full screening. In the meantime, PPBM was the security measure used most because it did not require the staffing and equipment acquisitions of the other acceptable policies. However, as Inspector General Kenneth M. Mead of the U.S. Department of Transportation testified, "...positive passenger bag match will not prevent a suicidal terrorist from blowing up an aircraft by putting a bomb in his baggage..." [102]. Given this, as airports began acquiring EDS/ETD machines and bomb-sniffing squads, PPBM was slowly phased out at most airports.

The question raised by this decision is whether the increased risk of allowing unaccompanied bags on planes outweighs the cost of maintaining PPBM. Critics of PPBM worry that bag-matching would cause system-wide delays and require additional staffing and equipment, costing the airlines money. Citing PPBM's inability to thwart suicidal terrorists as a major weakness, they argue that the threat posed by the few attack scenarios preventable by PPBM is not sufficiently large to justify this cost. On the other hand, supporters of PPBM argue that while EDS and ETD can be effective at screening bags, they are not perfect, and PPBM in conjunction with screening could help narrow the gap (See, e.g., [19]). Imagine, for instance, that a terrorist is planning to send an explosive in a checked suitcase. Under an EDS/ETD regime alone, he can arrive at the airport, check the suitcase and leave the

airport. If the EDS/ETD detects the suitcase, then the attack is averted, but it might be difficult, if not impossible, to apprehend the terrorist. Had PPBM been used in addition to EDS/ETD, he would first have to be willing to board the plane (and thus to commit suicide if EDS/ETD does not detect the bomb) *and* be willing to risk arrest (if the EDS/ETD successfully detect the explosive and officials apprehend the terrorist at the gate). Thus, PPBM can prevent attacks by non-suicidal terrorists who are unwilling to board the plane with the explosive, and it can also prevent attacks by suicidal yet otherwise risk-averse terrorists who are unwilling to be incarcerated. [41] note that "mission success is very important and leaders are in some ways risk-averse. Terrorists recognize that their power depends on perceptions of whether they are winning or losing;... martyrdom in a stymied mission lacks the appeal of dying in a spectacular, successful attack". Not all terrorists are suicidal, and furthermore, not all terrorists willing to risk their lives are also willing to risk imprisonment.

4.1.2 Moratorium on larger Postal Service airmail packages

The United States Postal Service (USPS) had been using commercial airlines for much of their airmail service since the 1920's. In 2000, 70% of the 2.5 billion revenue ton miles of airmail was shipped on passenger aircraft [147]. In the aftershock of the September 11 attacks, another change to aviation security procedures was the removal of USPS packages weighing more than one pound from the cargo holds of passenger planes [141]. (Prior to 9/11, USPS requested permission from its customers shipping larger packages to screen such packages if necessary [48].) It was believed that terrorists might try to ship explosives through the airmail system, and that one pound or more of explosives would be sufficient to down a passenger aircraft [67]. This new law forced the USPS to route larger packages on all-cargo airlines, such as Federal Express, resulting in a loss of revenue to the airlines and additional costs to the Postal Service[1]. Because the airlines were losing at least $250 million in revenue annually under the moratorium [7, 131], in May 2003, the Transportation Security Administration (TSA) began a pilot program at eleven airports where the larger packages would be permitted on passenger aircraft provided they were first screened by bomb-sniffing dogs. The intent of the pilot program was to test the bomb-sniffing dog policy in preparation for nationwide adoption [141]. However, not publicly addressed was the risk posed by attacks on airmail, the extent to which this risk can be mitigated by the bomb-sniffing dogs, and how this risk balances with the costs of the moratorium. Is the additional revenue to the airlines and the reduction in costs to the USPS sufficiently large to justify the increase in risk incurred if airmail packages were to be placed back on passenger aircraft?

[1] Although the Postal Service and FedEx had entered into an agreement in August 2001 where FedEx would carry some USPS airmail on their all-freight aircraft, this agreement included only mail on certain routes where it was economically advantageous. When packages were prohibited from passenger airlines, the USPS was forced to expand its agreement with FedEx.

4.1.3 A hypothetical policy on cargo

While airmail packages weighing only a pound are forbidden from passenger aircraft, significantly larger cargo shipments continue to be carried. In 2000, roughly 20% of the 12.2 billion revenue ton miles of air freight in the U.S. was shipped on passenger aircraft [147], and less than 5% of this was screened for explosives [48]. Although cargo screening is mandated in ATSA, the primary policy in use is the "Known Shipper" program, where cargo sent by companies with a history of working with the airlines is permitted on passenger aircraft, mostly unscreened. Any shipments by unknown shippers are sent on all-cargo carriers [147]. The concern is, therefore, that a terrorist having connections to the shipping department of a Known Shipper could possibly send explosives as cargo on passenger aircraft. This possibility was made apparent in 2003 when a man working in the shipping department of a company successfully shipped himself in a cargo package from New York to Texas [3, 82]. Though he was not a terrorist, and his particular package was not sent on a passenger aircraft but on an all-cargo carrier, this highlights the possibility that a Known Shipper might not always be able to control the actions of its employees. Terrorists might also forge shipping documentation and could take advantage of other loopholes [119].

Because of the difficulty in developing technology capable of screening cargo for explosives, however, laws requiring additional screening of cargo have stalled in Congress. Asa Hutchinson, Under Secretary for Border and Transportation Security, responded in 2003 to a proposed cargo security bill saying, "Only a small percentage of the nation's air cargo could be physically screened efficiently with available technology without significantly impeding the supply chain" [36]. And of the $30.4 billion Homeland Security 2004 funding bill, only $85 million was intended for air cargo screening [95].

The question we explore here is whether or not it is contradictory to allow unscreened cargo to be sent on passenger aircraft when significantly smaller airmail packages are forbidden from these same aircraft at most airports. Does the removal of airmail packages set a precedent for the case of cargo, and if not, under what conditions would the removal of cargo from passenger planes make sense?

4.2 A cost-benefit model

In this section, we develop a model that finds the threat of attack required to justify a particular security policy, and we discuss how to interpret these threat thresholds in the time horizon. We will later use this model to evaluate the three policies discussed above.

A common technique for calculating an indifference point between two opposing policies is to compute the expected value (or cost) of these policies and determine parameter values for which they are equal. We use this technique here as we compare the expected costs of instituting each of the above security measures to the expected costs of not instituting them, as a function of the cost of a terrorist attack that would have been prevented by the policy, and the likelihood of such an attack.

Our approach is similar to that in [124], who examine an "efficiency criterion" for a

security policy: how does the incremental benefit (risk reduction) of a particular policy compare to its incremental cost, or put another way, how much protection is the public willing to buy to achieve such a risk reduction? We phrase the question slightly differently, by asking what must be the risk reduction (measured as a decrease in attack likelihood) anticipated from a measure in order for society to be willing to pay for this measure?

Other work also emphasizes the likelihood of an attack. In quantitative risk assessment, in addition to understanding possible attack scenarios and their consequences, an equally important question is the likelihood of such scenarios occurring. While [13] and [53] suggest estimating the likelihood of terrorism scenarios based on the occurrence of similar events, we instead use the probability of attack as a decision threshold. Rather than knowing the exact likelihood of a particular attack, it is sufficient to know only whether the true likelihood is higher or lower than this threshold.

[33] perform a cost-benefit analysis of a counterterrorism policy by focusing on resource diversion rather than the likelihood of an attack. The authors compare several possible technologies to thwart attacks by MANPADS, Man Portable Air Defense Systems, on passenger aircraft. They compare costs of attacks having varying magnitudes with estimated costs of anti-MANPAD measures, and view these security costs as funds that would have to be diverted from other Department of Homeland Security initiatives. The decision of whether a particular measure is cost-effective or not depends on the measures that would be supplanted by it. This is a different framework than that to be presented in this chapter as it assumes a fixed Homeland Security budget and assesses measures in the context of that budget. We do not consider budgetary constraints, arguing instead that if the likelihood of a terrorist attack is sufficiently high, it is in society's best economic interest to implement counterattack measures, rather than to risk incurring the costs of an attack.

4.2.1 Parameters and constants

We consider a proposed policy that protects against a specific type of attack (for instance, PPBM protecting against IED's in unaccompanied checked luggage). If we choose not to implement the policy, then we assume a backup security plan (such as the original security policy in place), which has associated with it the following parameters:

- C_B: the annual cost of maintaining any backup security that is used in lieu of the proposed policy,
- p_B: the effectiveness of such backup security (i.e., the probability that this backup device can detect the attempted plot),
- C_A: the cost of a successful terrorist attack that might occur,
- r_{pre}: the risk per flight of an *attempted* attack in the absence of the proposed policy. We allow for the possibility that this attempt might be thwarted by the backup security system (which happens with probability p_B), but assume that if not detected, the explosive will detonate.

The parameters associated with the proposed policy are:

- C_P: the annual cost of maintaining the policy,
- F: the number of U.S. domestic flights annually, estimated at roughly 6 million per year,
- r_{post}: the probability per flight of an *attempted* attack after the policy has been implemented. This allows us to assume that there may be some deterrence effects of implementing the policy, such that the chance of an attempted attack decreases,
- p_P: the effectiveness of the policy (i.e., the probability that the policy can thwart an attempted attack).

Throughout this chapter, we will focus on the probability of an *attempted* attack rather than a successful attack. This allows us to isolate and vary the effectiveness of the backup security policy and explore, for instance, whether the proposed policy is still worthwhile even if the backup policy is reasonably good.

4.2.2 Model based on cost per flight

There are two logical models for determining whether or not to implement a security policy. The first is to estimate the expected cost of the security policy on a *per flight* basis and compare this to the expected cost per flight without the policy. The per flight cost is the annual cost of the policy divided by the number of flights, plus the cost of an attack that occurs on a flight weighted by the likelihood of such an attack. Using the parameters defined above, we calculate the expected value per flight of implementing a policy as follows:

$$\begin{aligned} E[\text{Value of policy per flight}] &= \text{Cost of not implementing policy} - \text{Cost of implementing policy} \\ &= [C_B/F + r_{pre}(1-p_B)C_A] - [C_P/F + r_{post}(1-p_P)C_A] \\ &= (r_{pre}(1-p_B) - r_{post}(1-p_P))C_A - \frac{C_P - C_B}{F}. \end{aligned} \tag{4.1}$$

In order for the policy to be worthwhile we must have $E[\text{value per flight}] \geq 0$, or

$$r_{pre} \geq \frac{C_P - C_B}{FC_A(1-p_B)} + r_{post}\frac{1-p_P}{1-p_B}. \tag{4.2}$$

Thus, if our threat of *attempt* per flight, absent the policy, is at least $\frac{C_P - C_B}{FC_A(1-p_B)} + r_{post}\frac{1-p_P}{1-p_B}$, then we should implement the policy, otherwise we should not.

4.2.3 A renewal model

We might also use a renewal method to estimate this threat threshold. This method uses the assumption that if the policy is not initially implemented, it will be implemented once an attack that might have been prevented by the policy occurs. Up until the first attack, there

is only the cost of the backup policy, but once an attack occurs, we incur both the cost of the attack and the cost of instituting this new security measure at this time and maintaining it throughout the future. It turns out that this method yields the same threshold as the cost per flight method.

To show this, we assume first that attack attempts absent the proposed policy follow a Poisson process of rate $r_{pre}F$ attempts per year. Given that the backup security measure has a chance p_B of thwarting the attempt, successful attacks therefore follow a Poisson process of rate $(1 - p_B)r_{pre}F$. In this case, the expected time, T, until the first successful attack takes place is $\frac{1}{(1-p_B)r_{pre}F}$. Until the first attack, we incur the cost of the backup policy, C_BT; at time T, we incur a cost of C_A due to the attack; and from time T until infinity, we incur an annual cost of C_P for implementing the policy after the attack, plus the cost of any subsequent attacks not prevented by the policy. Thus, the total expected cost if the policy is not implemented immediately is:

$$E[\text{cost of no policy}] = \frac{C_B}{(1 - p_B)r_{pre}F} + C_A + E[\text{Total Policy Cost after time } T]. \tag{4.3}$$

If, instead, the proposed policy is implemented, attempted attacks will occur at rate $r_{post}F$ per year, and successful attacks with rate $(1 - p_P)r_{post}F$ per year. In this case, up to time T (letting T be the same value as in the previous equation), we have the annual cost of the policy, C_P, incurred over T years. Over this time period, the expected number of successful attacks, given that the proposed policy is in place, is $(1-p_P)r_{post}FT = \frac{(1-p_P)r_{post}F}{(1-p_B)r_{pre}F}$. From time T until infinity, we incur the same cost of maintaining the policy that we had above. This gives us the total expected lifetime cost of implementing the policy:

$$\begin{aligned} E[\text{cost of policy}] &= \frac{C_P}{(1 - p_B)r_{pre}F} + \frac{(1 - p_P)r_{post}}{(1 - p_B)r_{pre}}C_A \\ &+ E[\text{Total Policy Cost after time } T]. \end{aligned} \tag{4.4}$$

If the cost of instituting the policy is less than that of not instituting it, then the policy has a positive expected value and should be implemented. This occurs when $r_{pre} \geq \frac{C_P - C_B}{FC_A(1-p_B)} + r_{post}\frac{1-p_P}{1-p_B}$, the same threshold we saw above using the cost per flight method.

This threshold can be further simplified for the three specific cases we consider here. Removing airmail or cargo from passenger aircraft, for instance, would completely eliminate the risk posed by such parcels. In this case, $p_P = 1$ (or equivalently, $r_{post} = 0$), and we have the threshold

$$r_{pre} \geq \frac{C_P - C_B}{FC_A(1 - p_B)}. \tag{4.5}$$

The case of PPBM is somewhat trickier. Some attempts to put explosives in luggage would be thwarted (those in which the terrorist leaves the premises and the bag is pulled by

PPBM) while others would not be (those in which the terrorist is willing to board the plane with the explosives *and* the EDS/ETD machines fail to detect the IED). As such, we define two types of attacks involving explosives placed in checked luggage:

- Type 1 attacks are those *preventable* by PPBM (attacks by nonsuicidal terrorists or by terrorists willing to commit suicide but not willing to risk arrest). These attacks will not occur if PPBM is instated and may or may not occur if it is not, depending on the performance of the EDS/ETD at detecting the explosive.

- Type 2 attacks are those that are *not preventable* by PPBM. These are attacks by suicidal terrorists who are willing to risk the possibility of arrest. These attacks could occur even if PPBM is used, again depending on the performance of the EDS/ETD at detecting the explosive.

Using the framework presented earlier, we can see that $r_{pre} = r_1 + r_2$, the risk per flight of an attempt on checked luggage, absent PPBM, is the overall risk per flight of Type 1 and Type 2 attempts (we assume that the chance of a Type 1 and Type 2 attempt occurring simultaneously on the same flight is zero). r_{post}, the chance of attempt in the presence of PPBM is then equal to r_2. Furthermore, given that a Type 2 attack is attempted, it can be stopped by the EDS/ETD machines with probability p_B. Thus, when the PPBM policy is in place (which assumes that EDS/ETD continue to be used), its chance of thwarting Type 2 attempts, p_P, is equal to p_B. Substituting into $r_{pre} \geq \frac{C_P - C_B}{FC_A(1-p_B)} + r_{post}\frac{1-p_P}{1-p_B}$, we get $r_1 \geq \frac{C_P - C_B}{FC_A(1-p_B)}$, which is similar to the expression found for the cases of airmail and cargo, where we consider the threat of only Type 1 attempts. The rate of Type 2 attempts (those not preventable by PPBM) is irrelevant for this model because the costs associated with such attacks are incurred regardless of the presence of PPBM.

Because we have eliminated r_{post} from the expression for the risk threshold, we will simplify notation by using r in place of r_{pre} for the remainder of the chapter, and using as the risk threshold the expression

$$r \geq \frac{C_P - C_B}{FC_A(1-p_B)}. \tag{4.6}$$

4.2.4 Interpretation

Many of the parameters above are unknown random variables, and for any realization, i, of the parameters (C_{P_i}, C_{B_i}, C_{A_i}, p_{B_i}), a different risk threshold r_i can be calculated. We rely, therefore, upon a summary statistic that might be useful in interpreting the model. For each scenario (C_{B_i},C_{P_i},C_{A_i},p_{B_i}), and risk level r, we let $d_i(r) = r(1-p_{B_i})C_{A_i} - \frac{C_{P_i} - C_{B_i}}{F}$ be the expected value of implementing the policy in that scenario. We would like to find a threshold value of r, to be called r^*, such that $E[d(r^*)] = 0$, where the expectation is taken

over all scenarios:

$$\begin{aligned} E[d(r^*)] &= 0, \\ E\left[r^*(1-p_B)C_A - \frac{C_P - C_B}{F}\right] &= 0, \\ r^* E\left[(1-p_B)C_A\right] - E\left[\frac{C_P - C_B}{F}\right] &= 0, \\ r^* &= \frac{E\left[\frac{C_P - C_B}{F}\right]}{E[(1-p_B)C_A]} \\ &= \frac{E[C_B] - E[C_P]}{FE[C_A]E[1-p_B]}. \end{aligned} \tag{4.7}$$

The last equation holds if we assume the random variables C_B, C_P, C_A, and p_B to be independent, which appears to be a reasonable assumption. If the true chance of attempt per flight is greater than r^*, then the expected value of the policy will outweigh the expected cost in the long run. We will implement a policy, even if the threat of attempt is small, if the ratio of policy costs to attack costs is relatively high. A pleasing feature to this threshold is that it depends on the distributions of the parameters only through the first moment.

Relating risk thresholds to the time to first attempt

The difficulty with using risk thresholds is that they are difficult to interpret. What is meant by a risk per flight of, say, 10^{-9}? When events occur so rarely, and have such large impacts when they do occur, it is often difficult to distinguish perceived risks, which might be emotionally charged, from true risks. So there is benefit in adjusting the above thresholds to an appropriate scale and metric.

One simple way is to translate a risk threshold into the time by which there will be a $x\%$ chance that at least one attempt will have occurred. We prove the following theorem:

Theorem 4.1. *For $T = \frac{\ln(1-x/100)}{\ln(1-r^*)}$ and $r^* = \frac{E[C_B]-E[C_P]}{FE[C_A]E[1-p_B]}$, the expected value of implementing the policy is equal to (or greater than) zero if and only if there is an $x\%$ (or greater) chance of an attempted attack within the next T flights (T/F years).*

Proof. We begin by proving that having an $x\%$ or greater chance of attempt within T flights at the true risk level, r_{true}, implies a non-negative expected value of implementing the policy:

$$\begin{aligned} P(\text{At least one attempt in first } T \text{ flights}) &\geq x/100 \\ \Rightarrow P(\text{No attempts in first } T \text{ flights}) &\leq 1 - x/100 \\ \Rightarrow (1 - r_{true})^T &\leq 1 - x/100 \\ \Rightarrow T \ln(1 - r_{true}) &\leq \ln(1 - x/100) \\ \Rightarrow \frac{\ln(1-x/100)}{\ln(1-r^*)} \ln(1 - r_{true}) &\leq \ln(1 - x/100) \end{aligned}$$

$$\begin{aligned} \Rightarrow r_{true} &\geq r^* \\ \Rightarrow E[d(r_{true})] &\geq E[d(r^*)] = 0. \end{aligned}$$

By reversing the above steps we find that if the expected value of the policy is non-negative at our current risk level r_{true}, then the chance of attempt within the next $T = \frac{\ln(1-x/100)}{\ln 1-r^*}$ flights is at least $x\%$. □

For example, if the expected value of implementing the policy is greater than or equal to zero, then there is a 10% chance of an attempt in the next $\frac{\ln(0.9)}{F\ln(1-r^*)}$ years, a 20% chance within the next $\frac{\ln(0.8)}{F\ln(1-r^*)}$ years, etc. Focusing our attention on a 50/50 likelihood threshold, if there is there is at least a 50% chance of an attempted attack within the next $\frac{\ln(0.5)}{F\ln(1-r^*)}$ years, then we should implement the policy.

4.3 Parameter Estimation

With the cost-benefit framework in place, we can now evaluate the three security measures mentioned earlier: Positive Passenger-Bag Match, the removal of larger Air Mail packages from commercial passenger aircraft, and the removal of cargo from commercial passenger aircraft, by estimating their corresponding parameters. The term r^* in Theorem 4.1 above depends only on the expected values of the parameters and not on their distribution. We will therefore discuss plausible distributions for each parameter and then find the time thresholds corresponding to setting each parameter equal to the expected value. To examine how the thresholds vary with the parameters, we will also evaluate thresholds using low and high settings for each parameter. We will consider any costs from society's, rather than the airlines', perspective (at least to a first-order approximation) because the ramifications of terrorist attacks affect society as a whole. Considering costs from the airlines' perspective alone would underestimate the true costs of an attack. For example, [120] argues that airlines and airports (who before September 11 were primarily responsible for security screening) were typically less willing to pay for security measures than the general public because "the benefits flowing to their organizations from tightened security did not justify the added costs. However, from a social point of view a tighter security regime would have been desirable". (The issue of who should pay for security measures is out of the scope of this thesis, and the interested reader can instead refer to [38, 59, 85]) We start our parameter estimation with the cost of an attack.

4.3.1 The cost of attack

The three policies we considered all protect against a similar type of attack, namely the placement of an explosive somewhere in the cargo hold of commercial passenger aircraft. Because the damage incurred in each type of attack is likely to be roughly the same, we assume an identical distribution of C_A across all three policies.

The natural tendency in estimating attack costs is now to use September 11 as a benchmark. The short-term direct costs of September 11 (loss of physical assets, rescue and cleanup) have been estimated to be roughly $27.2 billion [92], but the indirect costs, including economic repercussions, are believed to be much larger, and fundamentally difficult to estimate. For instance, the airline industry was the hardest hit due to a sharp reduction in demand for air travel following September 11. The ATA estimates that during the two years following the 9/11 attacks, the airline industry lost roughly $25 billion [7]. Second, the insurance and reinsurance industries suffered losses of at least $30B [92]. Businesses and victims' families have received an estimated $38.1 billion in compensation, both from the government and from private donations [29, 44]. Lastly, the general economic malaise, present prior to September 11, but accelerated by the attacks, caused many industries to suffer. To further complicate matters, some industries benefitted from the attacks, such as defense and security industries, but occasionally at the expense of the government and tax-payers.

However, the September 11 attacks were, hopefully, atypical. They were coordinated attacks involving four hijacked planes, as opposed to an isolated attack on a single plane that we consider in our model. The Air Transport Association has estimated that the cost to the airline industry of a future attack on aviation would be roughly $5 billion [7]. However, depending on where the explosive detonates (while the plane is empty and parked in the hangar, or while it is full of passengers parked near a crowded airport, for instance), the costs could be significantly different from this estimate. Not only might the physical damages vary, but the economic disruption, so prominent after the 9/11 attacks, could vary as well. For instance, a RAND Corporation study [33] estimated that the reaction to an attack could include a temporary shutdown of airspace, inflicting from $1.4 billion to $70.7 billion in lost revenues to the industry. While the latter figure corresponds to a one-month shutdown, an unlikely reaction to a single attack that we consider here, it highlights, nonetheless, this variability. (Furthermore, some suggest that even knowing that an attack was attempted, let alone successful, might damage the delicate financial state of the airline industry [165]. However, we focus here only on costs associated with successful attacks).

We assume the attack costs come from a triangular distribution ranging from $0.5B to $15B and peaking at $5B. This was chosen so that the mode of the distribution would be at $5B, the Air Transport Association's estimate. Such a distribution has a mean of $6.83 billion. For sensitivity analysis, we will also examine C_A equalling the extremes of the distribution, $0.5 billion and $15 billion.

4.3.2 Policy costs

Positive Passenger-Bag Match

In 1997, Barnett et al [21] conducted a two-week trial of Positive Passenger-Bag Match at Chicago's O'Hare Airport. Although PPBM had been in use for years on international flights, the study explored whether such bag-match would be feasible on domestic flights. They studied delays imposed on Chicago flights as a result of bag-match, how such delays would

be propagated throughout the entire air system, and what other costs might be incurred if PPBM were implemented. They found that the average PPBM delay per flight would be about 1 minute systemwide. Furthermore, the airlines involved in the trial estimated that if PPBM were to be imposed on a permanent basis, they would need additional staff and special equipment for comparing who had boarded the plane with who had checked luggage, keeping track of where in the plane the luggage had been stored and conducting bag-pulls. Lastly, delayed flights could possibly result in overnight stays for some passengers, to be paid for by the airlines. Their study predicted that the sum of these costs would be between 25 and 52 cents per passenger.

After September 11, 2001, many airlines and airports initially used PPBM to satisfy the ATSA mandate for checked luggage screening, since explosive detection systems and trace detectors were in short supply. At this time, [75] conducted a survey of major airlines to estimate the actual costs of PPBM, and found that the delays and costs actually realized were significantly lower than those originally predicted in [21]. While the original study predicted roughly one minute of delay per flight system-wide as a result of the bag-matching and pulling process, the updated survey found that the true delay was actually only 7 seconds per flight (ignoring delay propagation effects). Furthermore, it revealed that no additional staff were hired as a result of PPBM and that no additional equipment was acquired.

Using the same delay propagation multiplier, 1.2, as in [21], the average systemwide delay per flight caused by PPBM is estimated to be 8 seconds, which at $20 per minute (the estimate used by [21]) totals $2.80 per flight, or 2.75 cents per passenger (assuming 610 million passengers on 6 million flights annually, post-9/11). [21] found that 1 in 2700 passengers would miss their connecting flight due to the average PPBM-induced delay of 1 minute. 15% of these misconnections would result in overnight stays at the expense of the airline. Scaling by the revised average delay of 8 seconds, the revised cost of overnight stays is less than one cent per passenger. Thus, to have originating bag-match, the total cost would be roughly 3 cents per passenger, or $18.3 million per year. To include bag-match on connecting flights, we note that roughly one-third of passengers require connections, so a rough estimate of the cost of connecting *and* originating PPBM would be $24.4 million, or 33% higher than the estimate for originating bag-match alone. We use these two values as estimates of C_P for PPBM.

Removal of Air Mail Packages

Carrying airmail has long been a simple means for the airlines to obtain revenue with little marginal cost. The ATA estimates that the revenue lost by the airlines in losing the right to carry airmail packages is roughly $250 million per year [7], while other estimates extrapolated from quarterly losses are closer to $350 million [131]. There is also the additional cost to the U.S. Postal Service in having to send these packages on commercial freight carriers, such as FedEx. [118] estimates this is at most twice the cost of sending the mail by the airlines, although others argue that the Postal Service benefits from improved service provided by FedEx [115]. We assume C_P ranges according to a triangular distribution from $250 million to $500 million, peaking at $250 million, the expected value of which is $333.3 million, and

Airmail and bomb-sniffing dogs

For the case of airmail, the pilot policy at a few airports allows larger airmail packages to be shipped on passenger planes provided they are first screened by bomb-sniffing dogs. So p_B in the case of airmail is the probability that an explosive placed in an airmail package is detected by a bomb-sniffing dog.

It is widely believed that well-trained dogs can achieve 95% accuracy at detecting trace levels of explosives that they have been trained to detect [14, 34, 42, 43]. Secret Service dogs, for instance, are tested weekly, and are retired if they do not demonstrate greater than 90% accuracy during the test [14]; TSA appears to recertify annually [141]. Many argue that dogs are better than even the best explosives detection machines due to their ability to examine large volumes of packages quickly and accurately [134, 141]. They can be trained to recognize up to at least ten different scents without losing accuracy, and they can retain these scents in their memory for long periods of time without refresher training [14, 164]. Their accuracy also appears unaffected by the presence of other masking odors, even at strengths 10-100 times greater than the target odor [51]. Accounts of an appropriate duty cycle vary, with some scientists claiming dogs can work for up to two hours without a break under comfortable climatic conditions, provided they are trained to do so [52]. Under normal working conditions in a cool, dry environment, others (including Secretary of Transportation Norman Y. Mineta) say 40-60 minutes is more reasonable [34, 158].

Others raise concerns that bomb-sniffing dogs' performance can be variable, often unbeknownst to the handler, and is sensitive to a wide array of environmental factors. If the weather is hot and humid, they can lose focus in as little as 10-20 minutes [116, 134]. Says Susan F. Hallowell of the Transportation Security Administration, "The problem with canines is that they are like little children with IQs of 10. It's very hard to keep their attention" [4]. Dogs also appear to be highly sensitive to the behavior of their handler, who might accidentally cue the dog to respond a certain way [116]. Poor handling can reduce their effectiveness to as low as 60% [42, 43], and the use of explosive material that the dog has never encountered before, or wrapping the explosive in a material preventing the emission of odor vapors, would likely drop this effectiveness to near zero [134].

p_B taking on the value 0% corresponds to these latter cases, or to the pre-9/11 situation where airmail was carried on passenger planes without any backup screening. The values 0.75 and 0.95 represent the range of performance levels these dogs might exhibit depending on the explosive used, the trainer's behavior and environmental factors.

Cargo

Estimating backup security effectiveness for the case of cargo is somewhat trickier because currently, there is no backup security method in place apart from the Trusted Shipper program. This program does not actually screen cargo (except for a very small fraction of items), but is rather an agreement where companies sending cargo on passenger airlines certify that they are not sending dangerous items as cargo.

As early as 1999, recommendations were made to research technologies suitable for screen-

ing cargo, but nothing has yet been implemented. Proposals for cargo screening have included the use of bomb-sniffing dogs or EDS technology similar to that employed for checked luggage but modified to accommodate large cargo pallets. The U.S. General Accounting Office (now the Government Accountability Office) feels that dogs are the best choice for screening cargo because EDS technology can be slow to screen densely packed pallets and currently cannot accommodate large cargo [147]. ETD is even more time consuming and seems an unlikely option for cargo-screening [48]. Thus, we assume that bomb-sniffing dogs are currently the most likely option, and we adopt the same values for p_B as in the case of airmail.

4.3.4 Backup security costs

The final parameter to estimate is C_B, the cost of the baseline security measure used in lieu of the proposed policy of interest. For the case of PPBM, explosives detection systems and explosives trace detectors would continue to be used even if PPBM were instated, so we set $C_B = 0$ for this case.

For the prohibition on airmail and cargo, the backup policy considered is the use of bomb-sniffing dogs. Acquiring an untrained dog costs roughly $3,000, training it from $2,000 to $5,000, and it typically has a working life of 7-9 years [79, 134]. More significant, however, are the annual expenses, since each dog is assigned to a salaried handler, who typically takes the animal into his home and is compensated for additional costs and overtime work in caring for it. In the case of United States Secret Service dogs, this corresponds to a 6% salary increase for the Secret Service agent, plus two hours of daily overtime wages [134]. There are also medical and retraining costs. As such, $100,000 seems to be a reasonable estimate for the annual cost of a government explosives dog team ([116] indicates that private companies can charge up to $200,000). There are 429 commercial airports in the United States. While not all of these likely carry sufficient airmail or cargo packages to warrant a bomb-sniffing dog team used solely for this purpose (some airports might also use dogs for luggage screening and general surveillance of the airport), other airports may require multiple teams to screen all of the packages. If we suppose that one to two dog teams will be needed at each airport for screening airmail or cargo, then we arrive at an estimate of C_B between $42.9 million and $85.8 million, which will serve as our low and high estimates. We will assume C_B is uniformly distributed over this interval, and has expected value $E[C_B] = \$64.4$ million.

4.4 Comparing the cost-effectiveness of cargo-hold measures

We apply these parameter estimates to the model of Section 4.2 to determine the minimum threat level needed to adopt each of the three policies. For each policy, Table 4.1 shows the threshold probability of attempt per flight from Equation (4.6) and the threshold number of years obtained from Theorem 4.1 such that if there is at least a 50% chance of an attempted attack in this time, then the policy should be implemented. Both are evaluated with each parameter taking on its expected value as estimated in Section 4.3.

Policy	r^*	$E[d] > 0$ Threshold (years)
PPBM	$1.79\text{x}10^{-9}$	64.7
Airmail	$2.62\text{x}10^{-8}$	4.4
Cargo	$2.86\text{x}10^{-7}$	0.4

Table 4.1: Threshold probability of attempt per flight, r^*, required for each policy's expected value to be positive, and threshold time (in years) such that a 50% chance of attempt within this time causes the expected value to be positive.

We see that as long as there is at least a 50% chance within nearly 65 years that a terrorist will *try* to place an explosive in an unaccompanied suitcase (even if the EDS/ETD machines might detect it before it is loaded onto the plane), then implementing PPBM is cost-effective. By contrast, one would need to anticipate an attempt on airmail within the next 4.4 years, and an attempt on cargo within the next 0.4 years (less than five months) in order to justify the expense (primarily in revenue lost by the airlines) of removing them.

The results for PPBM are interesting because PPBM is a security measure that was allowed to slip away quietly on domestic flights. Yet, of the three policies considered, it may be the most cost-effective. While attempted attacks on unaccompanied checked luggage might be particularly rare in our current era of explosives detection technology, as long as such an attempt is more likely than not to be attempted within nearly 65 years, the costs of PPBM will be outweighed by its benefits. By contrast, one must feel an attempt on cargo is imminent before considering its prohibition from passenger planes. The case of airmail is less conclusive. If there is a greater than 50/50 chance of an attempted attack on airmail within the next 4.4 years (even if bomb-sniffing dogs might intercept the explosive), then the value of removing airmail exceeds the cost. It is difficult to know in this case whether our true risk is higher or lower than this threshold. There have been no attacks on airmail since September 11, 2001, but with the exception of a handful of trial airports, most airports were not allowed to load airmail packages onto passenger planes anyway. Prior to September 11, 2001 such an attack was committed by Ted Kaczynski, the Unabomber, in 1979 and threatened again in 1995.

In addition to considering thresholds corresponding to a 50% likelihood of attempt, we can let the value of x in Theorem 4.1 vary, as is shown in Figures 4-1 through 4-3, again with parameters set to their expected values. For each value on the x-axis, we must feel an attempt has probability x or higher of occurring within the corresponding number of years plotted on the curve for the policy to be cost-effective. For instance, selecting $x = 50$ corresponds to the 50/50 thresholds of Table 4.1. The thresholds shown on the curves are equivalent for any value of x: Selecting $x = 10$, we must feel there is at least a 10% chance of an attempt on checked luggage within approximately ten years, a 10% chance of attempt on airmail within eight months and a 10% chance of attempt on cargo within three weeks to be willing to accept the proposed policy. This is equivalent to selecting $x = 80$ and believing there is at least an 80% chance of an attempt on checked luggage within approximately 150

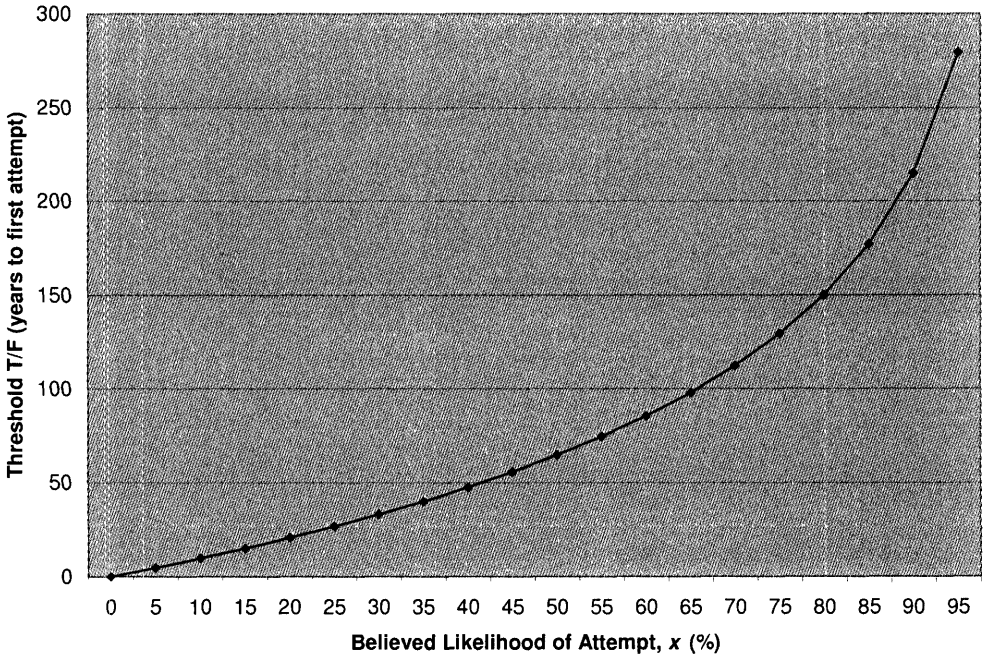

Figure 4-1: PPBM: Time threshold for which the expected value of the policy is positive, versus attempt likelihood, when parameters are taken at their mean estimates

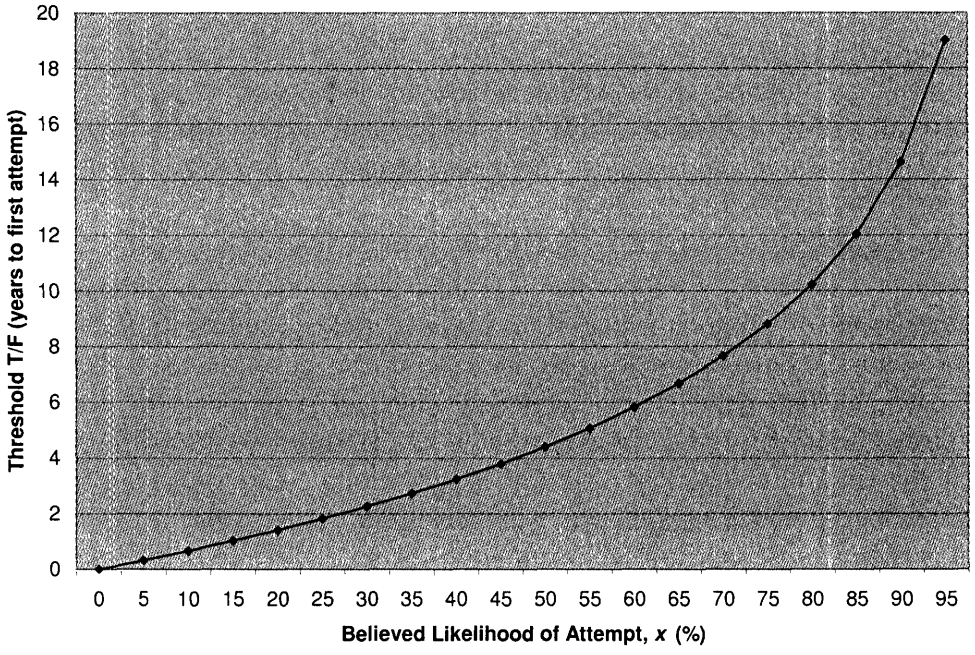

Figure 4-2: Airmail: Time threshold for which the expected value of the policy is positive, versus attempt likelihood, when parameters are taken at their mean estimates.

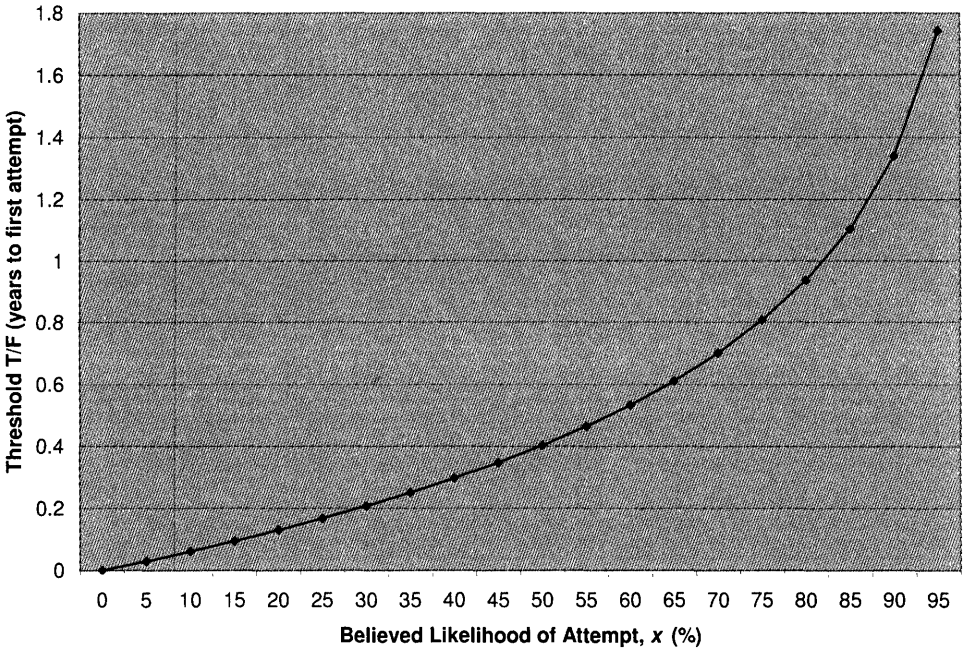

Figure 4-3: Cargo: Time threshold for which the expected value of the policy is positive, versus attempt likelihood, when parameters are taken at their mean estimates.

years, an 80% chance of attempt on airmail within 10.2 years and an 80% chance of attempt on cargo within eleven months to be willing to accept the proposed policy. We note that there is roughly an order of magnitude difference in time thresholds between PPBM and airmail, and between airmail and cargo, which is reasonable given the order of magnitude difference between the costs of the policies.

We can also explore how sensitive the thresholds are to variations in the parameter estimates. Figure 4-4 shows the threshold times until the first attempted attack under a 50% likelihood for the range of values considered for C_A, C_P and p_B under the PPBM policy. We see immediately that the parameters causing the largest change in the time threshold are the attack cost and the backup security effectiveness, where an increase in the attack cost from \$0.5 billion to \$15 billion causes the time threshold to increase (we are willing to implement PPBM even if the time until the first attempt is quite long), and where an increase in the backup security effectiveness decreases the time threshold (if our alternate security is quite good, then an attempt must be likely to occur soon for us to be willing to incur the costs of PPBM). We see time thresholds as little as one year, when the cost of an attack is only \$0.5 billion and the backup screening is 95% effective, and over 500 years when the cost of an attack is \$15 billion and the backup screening is poor. For the average attack cost of \$6.83 billion, we see that changing the effectiveness of the backup security causes the threshold to range from around ten years when PPBM is coupled with highly effective backup screening to over 200 years when the backup screening is poor.

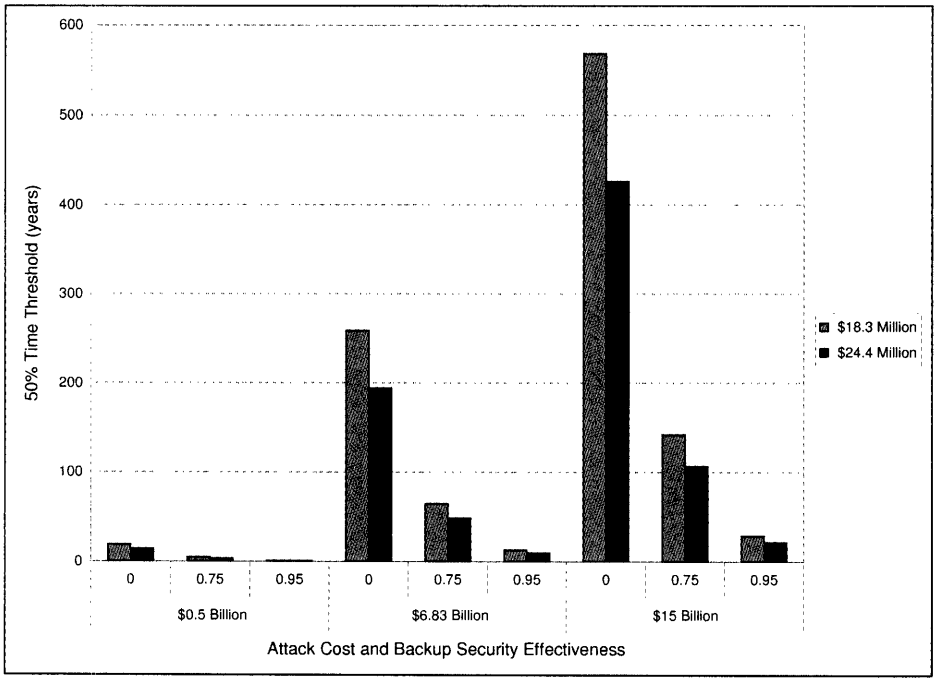

Figure 4-4: PPBM: 50% threshold time to first attempt by attack cost and backup security effectiveness, over policy costs of $18.3 million and $24.4 million.

Figures 4-5 and 4-6 show similar charts for the cases of airmail and cargo. For these policies, the threshold also depends on the cost C_B of the backup security measure, bomb-sniffing dogs, which is not shown these figures. However, because the range in which C_B was allowed to vary is quite small compared to the cost of either policy, variations in C_B were found not to affect the threshold time significantly. For the case of airmail, we see that the 50% probability threshold time until an attempted attack ranges from as little as 2 weeks (if the cost of an attack is very low, the policy cost is $500 million and the bomb-sniffing dogs are believed 95% effective) to as high as 41 years (for a high cost of attack, low policy cost and no backup security). Furthermore, at the mean attack cost of $6.83 billion, if there is no backup security in place (as was the case prior to September 11), then as long as there is a 50/50 chance of an attempt on airmail within 10-20 years (depending on C_P), then the removal of airmail is cost-effective.

In the case of cargo, even if an attack is believed to be very costly and there is no alternative security policy available, we still would need a 50% chance of attempt within approximately five years to justify the expense of removing cargo. If effective backup security is available or if the attack costs are not quite so high, this time threshold drops even lower. Thus, we still would require evidence of an imminent attack in order to remove cargo from passenger aircraft.

This emphasizes that despite the uncertainty in the parameters' true values, we can still draw important conclusions. Even if we determine that removing airmail from planes is cost-

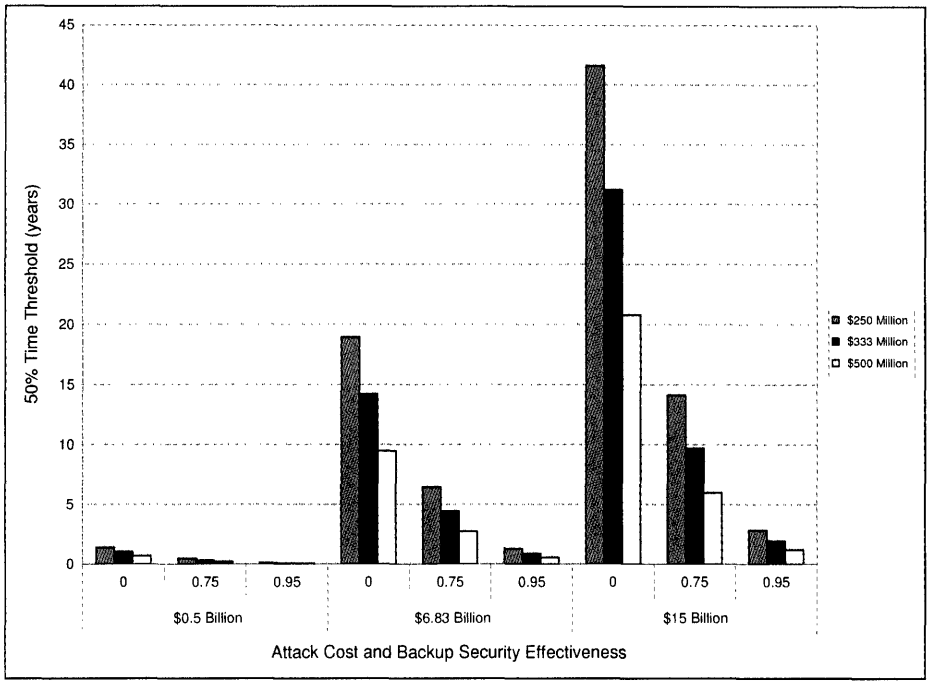

Figure 4-5: Airmail: 50% threshold time to first attempt by attack cost and backup security effectiveness, over policy costs of $250 million, $333 million and $500 million.

effective, this does *not* set a precedent for cargo, despite the similarity in the two policies and the type of attacks they prevent. The costs of removing cargo are roughly ten times those of removing airmail, so we require a significantly shorter period of time until the first attempted attack in order to justify cargo's removal. Similarly, unless the costs of a terrorist attack on checked luggage are quite small, continuing to use positive passenger-bag match in addition to explosives detection might be a cost-effective policy.

4.5 Diversion of threats

One criticism raised against this analysis is that it focuses on measures against seemingly small threats (such as matching luggage to passengers even when explosives detection technology is in use) when one could argue larger homeland security risks exist requiring immediate attention. Says, Charles V. Peña, director of defense policy studies at the Cato Institute, "You can only think of maybe a million things that a terrorist might do, and then you have to ask yourself if you're prepared to pay the costs of dealing with each and every one of them" [165].

This chapter shows that society should be prepared to pay the costs of dealing with even minute security threats *if* the likelihood of such threats is sufficiently high and the cost of addressing them sufficiently low. If a terrorist attack occurs, the costs incurred could be significantly larger than the costs of preventing such an attack. However, we have

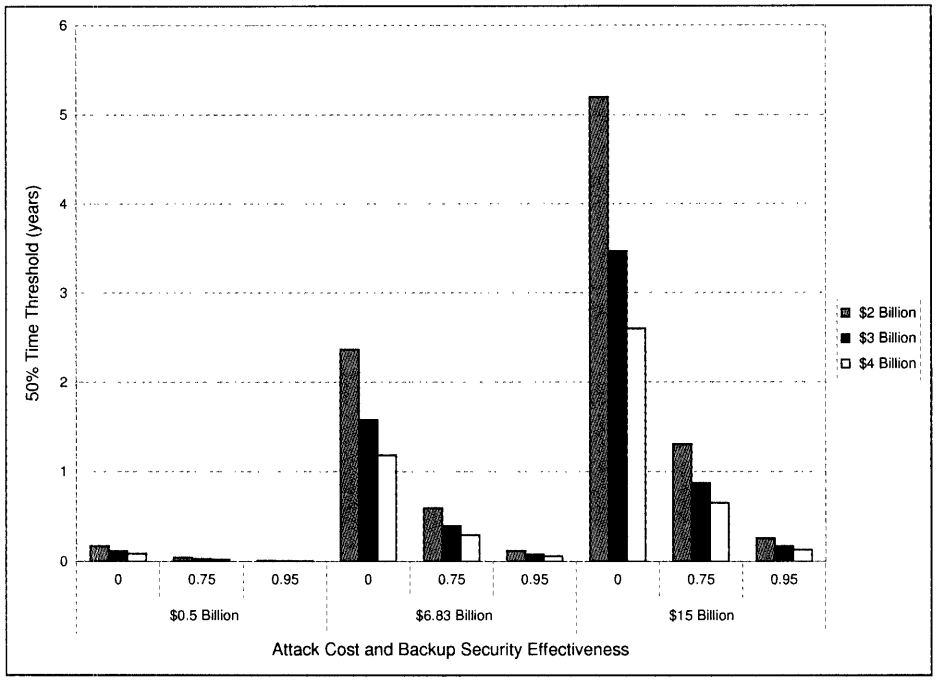

Figure 4-6: Cargo: 50% threshold time to first attempt by attack cost and backup security effectiveness, over policy costs of $2 billion, $3 billion and $4 billion.

argued in this chapter that mathematical analysis should be used to help identify those measures that are most cost-effective. "Without a comprehensive plan that incorporates a risk management approach, TSA and other federal decisionmakers cannot know whether resources are being deployed as effectively and efficiently as possible to reduce the risk and mitigate the consequences of a terrorist attack", United States General Accounting Office, 2002 [147].

Nevertheless, there is the possibility of threat *diversion*. We assumed in our model that r_{post}, the probability of an attempted attack if the proposed policy is implemented, is zero. For instance, if airmail packages are completely removed from aircraft, then there is no opportunity for an explosive to be placed in airmail and end up on a passenger plane. Yet terrorists, aware that this particular plot will no longer succeed, might then divert their attentions towards a less-hardened target. So while the likelihood of the particular type of attack considered has been driven to zero, the probability of a different attack occurring may increase.

However, we submit that while the post-intervention probability of attempt might not completely disappear, it will diminish, and this too can be captured in the model (by supposing a non-zero value of r_{post} in Equation (4.2). Because no small subset of security policies can guarantee to completely eliminate the threat of terrorism, to say that certain security measures are useless because they do not eliminate this threat is to suggest that *no* security measures should be implemented. According to terrorism experts at the RAND Corporation,

"Even if terrorists are not generally deterrable, specific terrorist *actions* may be deterrable even today. We know empirically that terrorists feel constraints, that they argue and plot amongst themselves, review and adapt strategies, worry about their perceived constituencies, and sometimes back away from tactics that seem to have gone too far" [41]. By securing obvious access points to aircraft, terrorists might be forced to explore riskier plots with lower net payoffs, and could, hence, be deterred.

4.6 Conclusion

This study has shown that just because a type of security policy might be efficient in one context, it is not necessarily efficient in all contexts. We saw that even when assuming a high rate of effectiveness of bomb-sniffing dogs, it might be more cost-effective to society to remove the airmail completely from passenger aircraft if the threat of attempt is sufficiently high. However, that bomb-sniffing dogs might not be an efficient strategy for the mail does not imply that they are inefficient for all types of packages. To the contrary, we saw that removing cargo completely from aircraft would probably never be a cost-effective option, and that bomb-sniffing dogs could provide a reasonable alternative. The airlines needn't fear that keeping airmail packages off passenger aircraft would set an expensive precedent for the case of cargo. Unless we are in imminent danger of suffering an attempt of this type, it is not cost-effective to remove cargo entirely from the planes, even if backup screening measures are unavailable.

We have also shown that in the upheaval following September 11 Positive Passenger-Bag Match slipped by the wayside despite the fact that its continuance would have required minimal cost (for originating PPBM, 3 pennies per customer) and would have provided a net expected benefit even under extremely low levels of risk. Its removal can be taken as a sign that quantitative analysis can help guide aviation security policies. We have provided a mathematical framework that can be used as a decision tool by those developing such policies.

Chapter 5

Dynamic Security Screener Allocation

While the earlier chapters of this thesis have focused on evaluating and selecting aviation security measures, another important problem is to improve the operational performance of a given security measure. Two commonly cited concerns with passenger screening at security checkpoints is the high level of manpower required and the delays experienced by passengers. Secretary of Transportation Norman Y. Mineta stated that "Passengers should not have to wait longer than 10 minutes in the security line" [158], yet reducing waiting times often requires an increase in staffing at the queue. Can operations research help reduce staffing levels and/or passenger waiting times? We explore this question in this chapter by examining techniques for the efficient allocation of security employees to different security checkpoints at an airport.

The direct benefit of efficient server allocation is reducing waiting times or staffing levels, but there may also be an indirect benefit on the quality of the security screening itself. It has been observed in some instances that human servers might work more quickly when they observe a long queue of waiting customers than if the line were short. While the opposite occasionally holds - servers become discouraged by the long lines and work slower - the emphasis placed by the TSA on efficient customer service as well as anecdotal evidence suggests that airport screeners might not be as thorough in examining x-ray images, verifying identification and conducting searches during peak periods as they are during quieter times of day. (Indeed, the author once observed a TSA screener wave passengers past the identification phase of security after a small line had formed). To the extent that long lines might negatively impact security, it is important to study how best to manage checkpoint queues.

The queueing problem studied here is motivated by techniques observed at San Francisco International Airport (SFO), where video cameras are focused on the different security checkpoints and project their images onto screens in a central control center. Transportation Security Administration (TSA) officials examine the monitors, looking primarily for security breaches. However, if they notice that the queue at one of the checkpoints is getting too long or suspect that it will grow quickly in the near future, they can decide to close a lane at a less busy checkpoint in the airport and transfer those screeners to a previously idle lane at a busier checkpoint. This decision of when to switch is made based on current queue lengths,

knowledge of future entry rates to the checkpoint according to scheduled flight departures, and employee experience. We use operations research techniques to find near-optimal anticipatory switching policies as a function of the system state, future entry rates and switching times. Though this research was inspired by the situation at SFO, airports across the U.S. could use the results of such work, and the underlying queueing problem is of general interest as well.

We find that while it is certainly important to vary server allocations over the day to accommodate changes in the customer arrival rate, the benefit of deviating *dynamically* from such schedules in response to stochastic fluctuations is minimal, unless the stochastic impulses are large and affect customers on an aggregate, rather than individual, level. In the following section, we describe the queueing decision problem and introduce parameters and terminology that are used throughout the chapter. Section 5.2 relates our problem to work already conducted on similar queue control models. In Section 5.3 we present five different formulations for this problem: one in which both arrivals to the queue and service times are deterministic, three in which the arrival pattern is altered in a deterministic or stochastic fashion and a formulation in which service times are stochastic. Section 5.4 addresses computational issues associated with these formulations and presents approximate dynamic programming techniques that are used in the analysis. To evaluate our models, we rely upon data provided by Boston Logan International Airport, which is described in Section 5.5. Section 5.6 presents the results obtained in this analysis and discusses the performance of dynamic allocations on the four models, and a discussion of the weaknesses of the models can be found in Section 5.7. We offer a few concluding remarks in Section 5.8.

5.1 Problem description

We consider two airport security checkpoints, A and B, either at two different terminals or within a same terminal but serving different aircraft departure gates. Arrival processes to these checkpoints are independent with time-varying, piecewise constant rates, $\lambda_A(t)$ and $\lambda_B(t)$. A fixed number, N, of screening teams (which we call *servers*), can be allocated to the two checkpoints, subject to the constraint that the number of screening teams at a given checkpoint cannot exceed the number of x-ray machines and metal detectors at that checkpoint (N_{max_A}, N_{max_B}). The question is how to find an allocation $\overline{(n_A, n_B)_t}$ at each decision epoch, t, such that $\overline{n_A(t)} + \overline{n_B(t)} = N$, $\overline{n_A(t)} \leq N_{max_A}$ and $\overline{n_B(t)} \leq N_{max_B}$ that minimizes the total time spent by customers waiting in line.

However, rather than determining an allocation schedule at the start of the day that cannot be changed, we allow this allocation to be determined on-line, at regularly spaced decision epochs of duration τ, as a function of the current system state (number of people in the respective queues, and current server allocation) and also knowledge of the future expected arrival rates. Thus, this is a *dynamic server allocation* problem where the goal is to determine optimally the conditions under which a server should be switched from one queue to another. If a queue loses one or more servers, this loss is experienced immediately, while the queue receiving the additional servers experiences a lag of θ minutes before the servers come

on duty, as they have to walk to the new checkpoint and activate the screening equipment. Hence, there is a tradeoff between responding to fluctuations in queue and arrival rate loads and maintaining sufficient service capacity. We note that because passengers arriving at the airport must wait in the security checkpoint line corresponding to their departure gate and are, therefore, not permitted to choose which queue they enter, load balancing can be achieved only by switching servers, rather than passengers, between queues.

To avoid confusion between people arriving at the security checkpoint to be screened and passengers arriving on incoming flights, we use the term "checkpoint entry rate" rather than "arrival rate" to refer to the rate at which people arrive at the security checkpoint to receive screening. We also refer to such people by the common queueing term "customers" rather than "passengers" since employees and vendors must also pass through the security checkpoints.

5.2 Literature review

A survey of Markov decision models used for controlling queue networks is found in [127]. Much of the literature pertaining to queue control addresses the optimal assignment of incoming *customers* (generally distinguished by classes having different arrival and service distributions) to parallel *servers*. See, for instance, [23, 61, 121]. The latter uses fluid models, which we will use in this chapter, to determine priorities of customer classes served at a same station. [55] and [64] consider the simultaneous allocation of servers and customers to service stations, and find that overall system performance improves as the individual facilities become more unbalanced in the number of allocated servers (the best policy is to assign one server to each station and then any extra servers to a single station). However, we note that a key difference in this problem is that after the servers have been allocated to stations, the optimal routing of customer classes to these stations can then be determined, whereas our problem assumes customer assignments to queues are fixed. Furthermore, they consider only static allocations that do not change in response to system evolution. All of these models assume that no switching costs or time are incurred if a server begins serving a new customer class and that the arrival and service distributions are stationary.

However, the situation at many airport security checkpoints is fundamentally different. First, customers must enter the queue corresponding to their departure gate, so any attempts to load-balance must be made on the service side, by switching screening teams between queues. There are also three additional characteristics that distinguish our work from the work in the literature: time-varying customer arrival rates, non-zero switching times, and decision epochs occurring at fixed intervals, rather than continuously or at "event epochs" marking a customer's entry or completion of service. While work in the literature incorporates one or two of these characteristics at a time, none appears to address all three simultaneously.

The simplest version of the dynamic server allocation problem assumes constant arrival rates and zero switching times when servers come on- or off-duty or are switched between queues. [105] is among the first papers in this category. They study a single queue having

exponential arrivals and services and develop thresholds to determine when additional servers should be brought on-duty and when they can be taken off, assuming a maximum number of servers available. They address the possibility of variable arrival rates by proposing that the problem be solved sequentially on segments of the time horizon over which the arrival rates are roughly constant, but their policy does not anticipate such variations. [57] studies a special two-queue, two-fixed-server problem where each queue is assigned to a single server. In addition, a third class of customers exists that are probabilistically routed to join one of the two other queues, and a third server is available that can choose which of the two queues to serve at any time. He shows the existence of switching curves that determine a threshold policy. [18, 27] explore parallel queues, having different service rates and holding costs, who are served by a single server. They demonstrate the optimality of the $c\mu$-rule, which assigns the server to the fastest, most expensive queue first, under an assumption of geometric service times and linear holding costs. [39] proves the existence of a stationary optimal policy for the dynamic allocation under heavy traffic conditions of M servers to N customer types where queues are not allowed to form (customers not assigned to a server are rejected from the system). [30] examine the dynamic control of a fluid queueing network with feedback and show that myopic optimization, minimizing the cost over the near future, yields a globally optimal solution over the entire time horizon. (In Section 5.3.1 we show that the myopic policy is not optimal when switching times are non-zero). [104] show that the optimal policy for the fluid limit model of a heavily congested network with deterministic routing is a good approximation for the optimal policy in the original model. They show that linear switching curves define the optimal policy for the fluid network and suggest that affine shifts of such curves (for instance, shifting the threshold so that the queue lengths stay close to their optimal mean values) might work well when translating the optimal fluid solution back into the discrete domain. [31] also find a relationship between an optimal fluid solution and the optimal discrete solution by showing that the optimal fluid solution can be used as an initial value for the dynamic programming value iteration algorithm on the stochastic model. [9] study servers in tandem and show that when service rates depend only on either the server or the customer class and not on both, all non-idling policies are optimal. The model studied in [126] most closely approaches our framework as they study the dynamic allocation of parallel servers to waiting jobs to minimize holding costs of customers in the system. They offer an example showing how the $c\mu$-rule, optimal in the fluid network, can be unstable in certain types of stochastic systems. In addition, they propose threshold policies for determining when to switch a server from one customer class to another. [94] shows heavy-traffic optimality of the generalized $c\mu$-rule when costs are convex. Most recently, [22] explores continuous control of queueing systems where service stations having one or more servers can be fractionally assigned to multiple customer classes.

There has also been extensive work on problems in which arrival distributions are again homogeneous but time is lost or a cost incurred whenever a server switches between customer classes. Such models are the subject of the survey paper found in [117]. Switching times typically render server allocation problems significantly more difficult, and policies, such as the $c\mu$-rule, that are known to be optimal in the case of zero switching times are generally

sub-optimal under non-zero switching times. (Indeed, [93] saw, in the case of M/M/1 queues in which the arrival and service rates can change in response to queue length changes, that placing costs on such changes causes *hysteresis*, where the changes occur less frequently than they would absent switching costs). Furthermore, only partial characterizations of optimal policies have been obtained for such problems. [56] looks at a system having fewer servers than queues and a deterministic transfer time between queues. At any time, only a single server can be assigned to a queue, and each queue might have a different service distribution. Assignment heuristics such as serve-longest-queue-first, first-come-first-serve and serve-shortest-processing-time-first are evaluated, and it is found that if the heuristics are modified to account for switching times, they perform better than if they ignore switching times. [65] and [83] examine a polling system having two queues with identical service distributions and holding costs and a single server. [65] proves that to minimize the sum of holding costs and setup charges, an optimal policy will be exhaustive: the queue being served must be depleted before the server can switch to the other queue. They and [83] propose a threshold policy as a likely optimal policy in this case. [84] uses dynamic programming to examine the two-queue, single server polling problem where the queues are not symmetric and provides switching thresholds in the limiting case as the queue with the smallest value of $c\mu$ gets long. [91] find polling policies for single-server, multiple-queue systems that stochastically minimize the number of customers in the system. They find, as above, that optimal policies are exhaustive and that for symmetric systems (where each queue has the same arrival, service and switching time distribution) the server will never switch queues if the system is empty, and when a switch occurs it will never switch to a queue having a stochastically smaller queue length. [45] partially characterize an optimal policy that favors "top-priority" queues and develop a heuristic policy for minimizing expected holding costs based on the $c\mu$ policy adapted for switching times. [10] approach the problem from the perspective of maximizing throughput rather than minimizing holding costs and use a linear program on the corresponding deterministic fluid model to determine the percentage allocation of each server to each customer class that achieves maximum throughput. They then construct round-robin allocation policies for the original stochastic model that achieve capacities arbitrarily close to the maximum capacity. However, because such policies can leave some queues unattended for long periods of time, they are unlikely to achieve minimum waiting times, which we seek here. The work of [24] offers an application for dynamic server allocations in the context of United States Postal Service offices, in which some employees are serving customers at the front desk while others are sorting mail in the back room of the post office. Similar to [105], if the queue of customers (or the total number of customers inside the store) grows large, back room employees can be brought out front to help serve customers. When the queue diminishes, they can return to sorting mail in the back. The objective is to minimize the total number of employees needed on staff, subject to a constraint on mean queue delay and a constraint requiring a minimum time average number of workers in the back room to ensure that all of the mail is sorted. Switching times are incurred on switches from the front room to the back room. The authors assume the existence of a threshold policy to determine when employees should be switched back and forth, and they develop

a heuristic to determine values of these thresholds for all possible system states, subject to constraints that switches from the back room to the front room are not allowed to occur prior to the formation of a queue and switches from the front room to the back room cannot occur if a queue remains, and must occur immediately following either an arrival or a departure from the customer queue.

The next class of server allocation problems features time-varying arrival processes, but no switching times or costs. [72] study a single queue having sinusoidal arrival rates over time and explore how to determine time-varying staffing levels, but do not address how to change the staffing level adaptively in response to realized arrivals, nor do they study tradeoffs when two time-varying queues must vie for a fixed number of servers. [98] examine server allocation amongst multiple queues in the context of airport customs procedures where customs officials can be switched (with no switching time) between arriving customs posts and departing customs posts subject to level of service requirements and limitations on the number of customers that can be re-queued when a post closes during a switch. They consider this problem on a tactical rather than operational level to construct a feasible daily schedule that cannot be changed dynamically and that is then used to construct weekly employee work schedules. The work examining *dynamic* allocation of servers in the context of nonhomogeneous arrivals with no switching times or costs appears to begin with [155] who explores the fluid flow model using differential equations to solve for the optimal continuous service rates, requiring an assumption of differentiability in the arrival and service process. To maintain differentiability of the service process, allocations of servers to classes are kept proportional to queue lengths (which are differentiable) and can be fractional. [130] estimates the average waiting time under dynamic telephone operator staffing in a single queue environment with abandonment and retrials where the staffing levels are estimated based on half-hourly load forecasts. He assumes, similar to [24] and [105], that operators not staffing the phones can engage in other lower-priority work until they are brought back on-duty again. A similar assumption is made in [163] who study a telephone call center where the goal is to immediately answer all calls. Changes in the staffing level are made based on the number of calls currently in progress and an estimate of the number of calls that will arrive and remain in service in the near future, which is similar to the anticipatory assumption we make here.

It is the final category, covering nonhomogeneous arrivals *and* non-zero switching times, that sets the work in this chapter apart from that in the literature. As we described earlier, problems involving non-zero switching times are typically more difficult than their zero switching time counterparts, and incorporating nonstationarity via the arrival process renders the problem even more intractable. As such, very little research in this area could be found in the literature. [50, 113, 114] study the problem in a general framework of M heterogeneous queues (whereas we consider only two queues having identical service distributions, switching times and holding costs, but different arrival distributions) and a total of N servers. However, in their framework, switches can take place only one server at a time (whereas we allow multiple servers to be switched at a time) and can occur almost continuously, whenever an arrival or service is completed as opposed to at regularly spaced

decision epochs that are independent of arrivals, services and switches, as in our model. Their focus on event epochs under memoryless distributions allows the use of Markov decision processes which can model state changes as simple one-step transitions. In our model, such a simplification is not possible because from one decision epoch to the next, there can be a sequence of many arrivals and departures. However, the most salient difference is that although they develop heuristics that can adapt to time-varying arrival rates (by assuming piecewise constant arrival rates and running the heuristics sequentially over the time intervals), these heuristics, and the dynamic program to which these heuristics are compared, do not use knowledge of *future* arrival rate shifts in determining the current period's allocation. Thus, their framework is actually closer to a constant arrival rate framework that is applied on a few time-varying test cases. This allows them to rely upon steady-state behavior and assume stationary policies, which we are unable to do here.

5.3 Problem formulations

We present now the optimization model for this problem. Our objective function is to minimize the total waiting time incurred by all customers who pass through the system. Because allocating additional servers to reduce Queue A, for instance, could cause Queue B to increase, the problem is therefore to determine an allocation that minimizes the *total* waiting time over the two queues. The total waiting time spent by customers in a queue is the integral of the instantaneous queue length over time, or the area between the cumulative entries and cumulative services curves, as shown in Figure 5-1.

Minimizing the total waiting time does not address the *variability* in waiting times, however. Under a minimum waiting time allocation, some customers might actually have to wait a very long time before service, even if most customers are served relatively quickly. Other possible objective functions could be to minimize the variance in waiting times between the two queues, or to minimize the fraction of passengers that must wait longer than, say, ten minutes. Nonetheless, much of the work in the literature uses holding cost rates that are linear in queue lengths (such as waiting time). As this work is among the first in the category of queueing control problems involving time-varying arrival rates and switching times, we focus on total waiting times, recognizing that other performance measures might be more appropriate and could be the focus of future research.

One simple allocation, which we call a *fixed allocation*, is to allocate $\overline{(n_A, n_B)}$ servers to Queues A and B at the start of the day and to keep this same allocation throughout the day. If $Q_i(t, n_i)$ is the instantaneous queue length at time t at Queue i having n_i servers on-duty, then the best allocation $\overline{(n_A, n_B)}$ is the one achieving the minimum in:

$$\min_{(n_A, n_B)} E\left[\int_0^T Q_A(t, n_A) + Q_B(t, n_B)dt\right], \tag{5.1}$$

such that $\max(N - N_{max_B}, 0) \leq n_A \leq \min(N_{max_A}, N)$ and $n_A + n_B = N$, and where the expectation is taken with respect to the stochastic entry and service times.

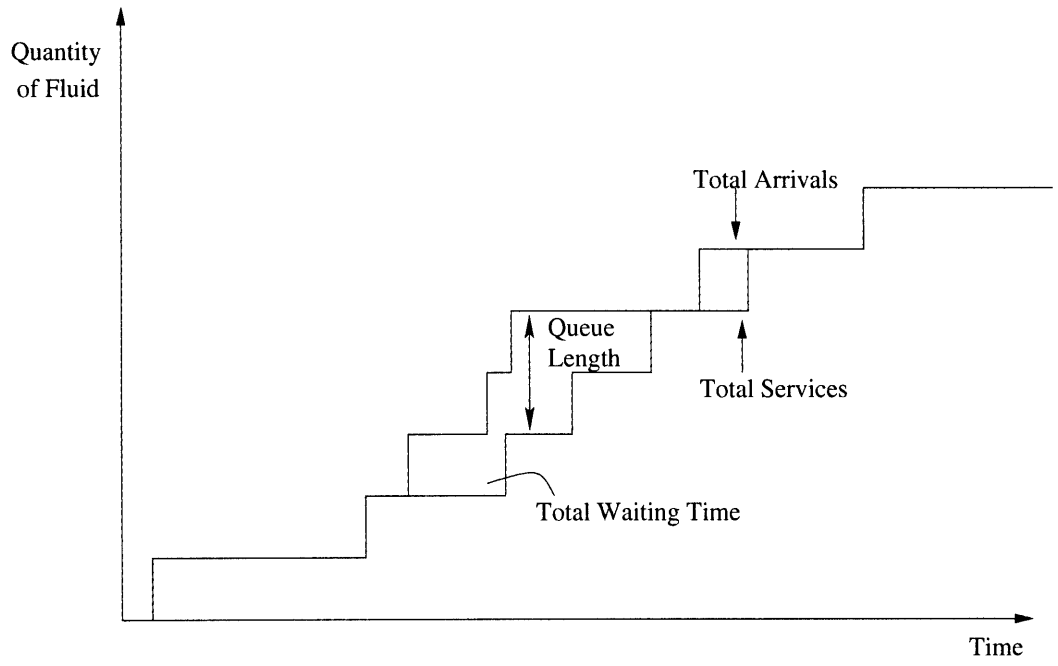

Figure 5-1: Cumulative entries (upper curve) and services (lower curve) in a queue having stochastic entry and service times. The instantaneous queue length is the vertical distance between the cumulative entry curve and the cumulative service curve. The total waiting time in a queue over a period is the integral of the instantaneous queue length over that period, represented as the area between the two curves.

This fixed allocation is quite restrictive, however, in that the allocation, once determined, is never allowed to change. Indeed, more realistic would be to use a *schedule allocation* where, at each time period, the screeners can be shifted between the queues based on expectations of customer entry rates at the queues. For instance, if Queue A tends to be quite busy in the morning while Queue B tends to be busy in the afternoon, it might make sense to allocate more servers to Queue A in the morning and then switch some over to Queue B in the afternoon. A schedule allocation changes only according to the time of day and not the particular stochastic evolution of the system, and can thus be determined at the start of the day for all decision epochs i. That is, we look for a pair of vectors, $\overline{(\vec{n}_A, \vec{n}_B)}$, (where $\overline{(\vec{n}_A(i), \vec{n}_B(i))}$ is the allocation during decision epoch i), achieving the minimum in

$$\min_{(\vec{n}_A, \vec{n}_B)} E\left[\sum_i \int_{t_i}^{t_i+\tau} Q_A(t, \vec{n}_A) + Q_B(t, \vec{n}_B) dt\right], \tag{5.2}$$

with $\max(N - N_{max_B}, 0) \leq \vec{n}_A(i) \leq \min(N_{max_A}, N)$ and $\vec{n}_A(i) + \vec{n}_B(i) = N, \forall i$. A schedule allocation is the one that typically comes to mind in the context of staff scheduling: it designates, for each decision epoch, how many employees are to be assigned to each post.

However, we are interested in a *dynamic* allocation that depends not only on the time period and *expected* entry rates (as in a schedule allocation), but that also can be modified as the day progresses based on the *actual* state of the system that arises due to the system's

Time	Fixed Allocation	Schedule Allocation	Q_A	Q_B	(N_A, N_B)	Dynamic Allocation
3:00	(7,3)	(7,3)	10	10	(7,3)	(7,3)
			0	30	(7,3)	(6,4)
					...	
3:30	(7,3)	(7,3)	0	30	(7,3)	(7,3)
			10	10	(6,4)	(7,3)
					...	
4:00	(7,3)	(8,2)	0	60	(7,3)	(6,4)
			0	30	(7,3)	(8,2)
					...	
...	...	...			...	

Table 5.1: Sample solutions under fixed, schedule and dynamic allocations. The fixed allocation cannot change during the day. The schedule allocation is determined in advance for each time period, regardless of how the system evolves stochastically, while the dynamic allocation can vary depending on the time and the system state (number in queue and initial allocation).

stochasticity. The optimal decision at time t depends on the current queue lengths, Q_A and Q_B, and current server allocation (N_A, N_B), as well as on the expected future entry rates $\lambda_A(t)$ and $\lambda_B(t)$, assumed to be known in advance. (For queues with general entry and service distributions, the system state also includes the values s_A and s_B, the time since the last customer entry to queues A and B, respectively, and the vectors $\vec{v_A}$ and $\vec{v_B}$ of time already spent in service by each customer in service at queues A and B.)

To summarize, a fixed allocation is determined at the start of the day and remains constant over all epochs. A schedule allocation is determined at the start of the day but provides a different allocation for each decision epoch. And a dynamic allocation is determined online, varying depending on the time of day, queue lengths and current server allocation. Table 5.1 illustrates how the three types of allocations (fixed, schedule and dynamic) can differ.

To find the optimal dynamic allocation, we use a dynamic programming formulation (DP). We define the following:

- $S = (Q_A, Q_B, (N_A, N_B), s_A, s_B, \vec{v_A}, \vec{v_B})$, the system state,
- $W_t(S)$, the expected "wait-to-go", or total waiting time incurred by all customers in the system at time t or who will enter the system from time t onward to the end of the day (time T), under an optimal allocation, starting in state S at time t,
- $w_t(S, (n_A, n_B), \theta)$ the waiting time incurred over the current decision period starting in state S at time t and selecting allocation (n_A, n_B), when switches require θ time units to be completed,

- $\hat{S}_{t+\tau}(S,(n_A,n_B),\theta)$, the state of the system at the end of the current decision epoch (time $t+\tau$) when starting at time t in state S, choosing allocation (n_A,n_B) and when switches require θ time units.

Starting with an empty system, S_0, the optimal dynamic allocation is the choice of $\overline{(n_A,n_B)}$ for each state S that minimizes the expected wait-to-go $W_0(S_0)$, where $W_t(S)$ is given by:

$$\begin{aligned} W_T(.) &= 0 \\ W_t(S) &= \min_{(n_A,n_B)} E\left[w_t\left(S,(n_A,n_B),\theta\right) + W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta)\right)\right], \forall t < T, \end{aligned} \tag{5.3}$$

where the expectation is taken with respect to the entry and service distributions and where $\max(N - N_{max_B}, 0) \leq n_A \leq \min(N_{max_A}, N)$ and $n_A + n_B = N$. The first term is the wait experienced over the current decision epoch by every customer already in the system or entering the system. Given the initial queue lengths and server allocation at time t, the new allocation chosen, the switching time θ and the stochastic of entries and services over the epoch $(t, t+\tau)$, one can compute the state of the system at time $t+\tau$, $\hat{S}_{t+\tau}$. The second term in (5.3) is then just the expected wait-to-go from time $t+\tau$ onwards, starting in state $\hat{S}_{t+\tau}$.

With all of the possible combinations of queue lengths, allocations and residual entry and service time vectors, s and $\vec{v}$, there is a large set of possible states at each decision epoch. The formulation in (5.3) is rendered even more complicated by the large number of possible trajectories of the state space (different sequences of inter-entry and service times) that must be considered in order to compute the expected values w_t and $W_{t+\tau}$.

To get around these difficulties, we use *fluid models*, in which customers are not considered to be discrete entities but rather as a fluid, entering and departing at continuous rates (refer to [108] for an introduction to fluid queues). While it is not completely accurate to replace discrete entities with a fluid, airport security checkpoint queues typically see a large volume of customers entering and departing as a flow, such that distinguishing individual passengers becomes less important. We no longer need to keep track of the residual inter-entry times, s_A and s_B, and the vectors $\vec{v_A}$ and $\vec{v_B}$, of residual service times for each customer in service, which greatly simplifies the model.

We begin by considering a *deterministic* fluid model in which not only are the entries and services of customers continuous, but the rates are deterministic (yet still time-varying). Temporarily considering only the deterministic case will allow us to avoid computing each possible stochastic trajectory of the system, as required in (5.3). After analyzing the deterministic case, we will introduce some uncertainty into the fluid framework, first by considering deterministic and then stochastic disruptions to the entry process, and then by considering a stochastic service process. For each of these variants, we find the three types of allocations described above:

- a best *fixed allocation*, where $\overline{(n_A,n_B)}$ is determined at the start of the day and is not

allowed to change,

- a *schedule allocation*, where a time-varying schedule is determined at the start of the day that indicates the best allocation for each time period,

- a *dynamic allocation* that is determined on-line, depending not only on time but also on the system state arrived at by the stochastic trajectory.

In the following sections, we discuss in greater detail the formulations for the deterministic and stochastic fluid models.

5.3.1 Deterministic fluid model

In our deterministic fluid model, the entry rates to Queues A and B, $\lambda_A(t)$ and $\lambda_B(t)$, are assumed piecewise constant, with changes coinciding with decision epochs. The service rate per server is constant at μ customers per minute, yielding a total service capacity at Queues A and B of $N_A\mu$ and $N_B\mu$, respectively, depending on the allocation (N_A, N_B) being used at the time. As discussed in [108] and shown in Figure 5-2, if the service capacity at a queue exceeds the entry rate when no queue has formed, then fluid is processed linearly at the entry rate (fluid cannot be processed faster than it appears at the queue). Because it is processed as soon as it enters, no queue is formed in this case, and the total waiting time remains zero. On the other hand, if the entry rate is greater than the total service rate, then a queue grows linearly with time at a rate equal to their difference. Whenever a queue has formed, fluid is processed at the maximum service rate. If this maximum service rate exceeds the entry rate, then the queue is depleted linearly with time at a rate equal to their difference, otherwise the queue continues to grow. The instantaneous queue length is again the difference between the cumulative entries and cumulative services curves, and we wish to find an allocation of servers that minimizes the total waiting time, which is the integral of the queue length over time, or the area between the two curves.

Obtaining a best fixed allocation is straightforward: for every feasible (n_A, n_B) pair, we calculate the total waiting time that will be incurred in the system over the day as a result of the allocation, then select the allocation yielding the smallest waiting time. The small number of possible allocations (there are at most $N + 1$) and the deterministic nature cause this to be a fast and simple calculation.

To find the best *schedule allocation*, where the allocation can change at each time period, we have to consider the future effects of each period's allocation throughout the day. In other words, at the start of each half-hour block, we find the allocation (n_A, n_B) that minimizes not just the short-term waiting time over that period but the total waiting time incurred over the remainder of the time horizon as a result of this and future optimal allocations. However, because of the deterministic nature of this model, an allocation made at a particular decision epoch causes the system to transition to exactly one possible state at the next epoch. These transitions can be predicted in advance, and as such, the time-dependent schedule allocation is actually equivalent to a state-and-time-dependent dynamic allocation. Therefore, we use

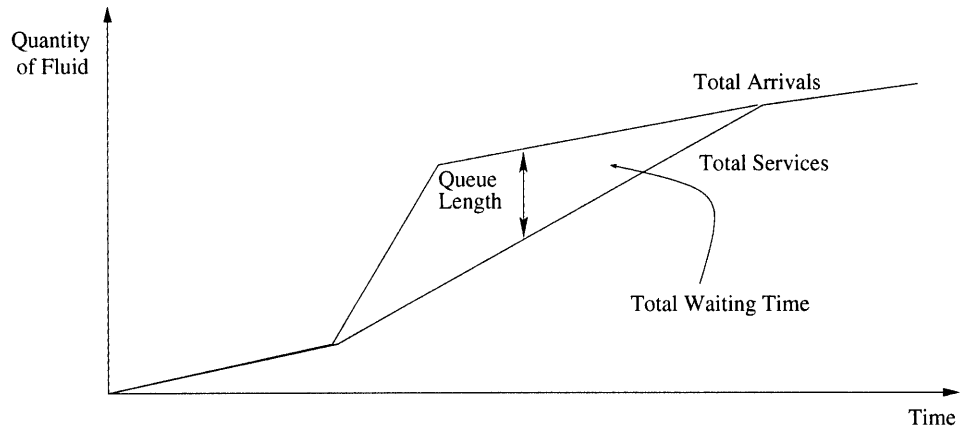

Figure 5-2: Cumulative entries (upper curve) and services (lower curve) in a fluid queue with piecewise constant entry rates. The instantaneous queue length is the vertical distance between the cumulative entry curve and the cumulative service curve. The total waiting time in a queue over a period is the integral of the instantaneous queue length over that period, represented as the area between the two curves. If the service capacity exceeds the entry rate before a queue has formed, then fluid is processed immediately as it enters and no queue forms. If the entry rate exceeds the service rate, then a queue develops that can be reduced only when the entry rate falls below the service capacity.

the dynamic programming framework given in (5.3), where the expectation is replaced by the true value.

To solve this, we must derive expressions for the waiting time function and the system state evolution. For any allocation (n_A, n_B) chosen to succeed a current allocation (N_A, N_B), if Queue i loses one or more servers (i.e. the new allocation has $n_i \leq N_i$), then this loss is felt immediately and is not subject to the switching time θ. So, for any allocation such that $n_i \leq N_i$, we can have one of the three cases shown in Figure 5-3. In the first, the initial queue is equal to zero, and the service rate $n_i\mu$ exceeds the entry rate $\lambda_i(t)$. Arriving fluid is processed immediately, so no queue is formed and no waiting time is incurred over the decision period $(t, t+\tau)$. At the end of the period, the new queue length, Q_i', is equal to 0, and the current period waiting time, w_i, equals 0. In the second diagram, we start with an initial $Q_i > 0$ but the maximum service capacity is high enough to work off the queue and all new entries to the system before the end of the decision period (i.e., $Q_i + \lambda_i(t)\tau \leq n_i\mu\tau$), leaving $Q_i' = 0$. Because the queue will be completely depleted at time $t = \frac{Q_i}{n_i\mu - \lambda_i(t)}$, the waiting time, w_i, is equal to the area of the triangle, $\frac{Q_i}{2}\frac{Q_i}{n_i\mu - \lambda_i(t)}$. In the final case, we have the initial queue length $Q_i \geq 0$, and the n_i servers on duty are unable to work off the queue and the new entering customers over the period (i.e. $Q_i + \lambda_i(t)\tau > n_i\mu\tau$). In this case, the end queue length, $Q_i' = Q_i + \tau(\lambda_i(t) - n_i\mu)$, and the waiting time incurred is the area of the trapezoid formed by the two curves, or $w_i = \frac{Q_i + Q_i'}{2}\tau$.

If a queue gains one or more servers ($n_i > N_i$), then it must wait a time of θ before

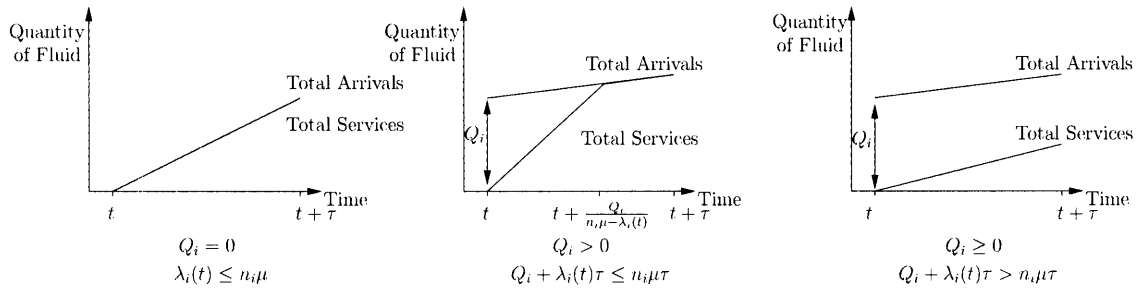

Figure 5-3: Three possible fluid system evolutions when the number of servers at a queue is decreased from N_i to n_i. In the first, the system starts empty and entries occur at a rate $\lambda_i(t)$ smaller than the new service capacity, $n_i\mu$. In the second, a queue exists at the start of the period, but the service capacity is large enough to deplete the queue by time $t + \frac{Q}{n_i\mu - \lambda_i(t)}$. In the third, the service capacity is insufficient to deplete the queue by the end of the period.

those servers arrive on-duty. For the first θ minutes of the period, therefore, there are the original N_i servers on duty (with a maximum service capacity of $N_i\mu$), and then after time θ, that number increases to n_i and the maximum service capacity to $n_i\mu$. The possible cases are shown in Figure 5-4. In the first, the initial Q_i is empty, and the service capacity of the initial allocation of servers, $N_i\mu$, meets or exceeds the entry rate $\lambda_i(t)$. Because the arrival of additional servers will only increase the service capacity, no queue forms over the entire period ($Q_i' = 0$), and, therefore, no waiting time is incurred ($w_i = 0$). In the second diagram, $Q_i + \lambda_i(t)\theta \leq N_i\mu\theta$, so the queue is depleted at time $t + \frac{Q_i}{N_i\mu - \lambda_i(t)}$, before the arrival of additional servers at time $t+\theta$. Again, $Q_i' = 0$, and we have $w_i = \frac{Q_i}{2}\frac{Q_i}{N_i\mu - \lambda_i(t)}$. In the third case of Figure 5-4, the queue is not depleted before the arrival of the additional servers, but it is depleted before the end of the period (i.e., $Q_i + \lambda_i(t)\theta > N_i\mu\theta$, but $Q_i + \lambda_i(t)\tau \leq N_i\mu\theta + n_i\mu(\tau - \theta)$). So $Q_i' = 0$ and $w_i = \frac{Q_i + [Q_i + \theta(\lambda_i(t) - N_i\mu)]}{2}\theta + \left(\frac{Q_i + \theta(\lambda_i(t) - N_i\mu)}{2}\right)\left(\frac{Q_i + \theta(\lambda_i(t) - N_i\mu)}{n_i\mu - \lambda_i(t)}\right)$. The last case is where we start with an initial $Q_i \geq 0$ and see that even once the new servers have arrived on-duty, there is still not enough capacity to deplete the queue before the next epoch ($Q_i + \lambda_i(t)\tau > N_i\mu\theta + n_i\mu(\tau - \theta)$). In this case, the queue length at the end of the period is given by $Q_i' = Q_i + \lambda_i(t)\tau - \mu(N_i\theta + n_i(\tau - \theta))$. The total waiting time w_i is the sum of the areas between the curves from t to $t + \theta$ and from $t + \theta$ to $t + \tau$: $\frac{Q_i + [Q_i + \theta(\lambda_i(t) - N_i\mu)]}{2}\theta + \frac{[Q_i + \theta(\lambda_i(t) - N_i\mu)] + Q_i'}{2}(\tau - \theta)$.

The expressions for Q_i' and w_i are summarized below. Using these expressions, we solve (5.3) to find the optimal schedule allocation for the deterministic fluid model, where $S =$

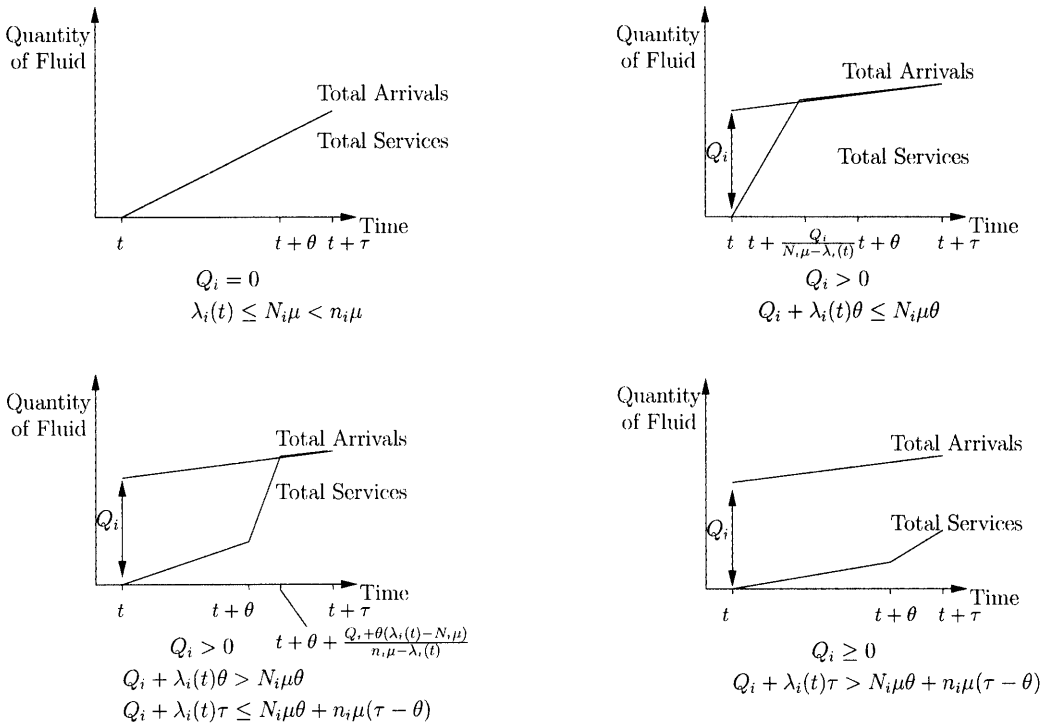

Figure 5-4: Four possible fluid system evolutions when the number of servers at a queue is increased from N_i to n_i. In the first, the system starts empty and entries occur at a rate smaller than both the initial and the new service capacities, $N_i\mu$ and $n_i\mu$. In the second, a queue exists at the start of the period, but the original servers are able to deplete the queue before the additional servers come on duty at time $t + \theta$ ($\frac{Q}{N_i\mu - \lambda_i(t)} < \theta$). In the third, the queue is not depleted before the new servers arrive, but the additional service capacity after time $t + \theta$ is sufficient to deplete the queue before the end of the period. In the last diagram, the service capacity is insufficient to deplete the queue by the end of the period, even with the arrival of additional servers at time $t + \theta$.

(Q_A, Q_B, N_A, N_B), $w_t = w_{1,t} + w_{2,t}$, and $\hat{S} = (Q'_A, Q'_B, (n_A, n_B))$.

$$Q'_i = \begin{cases} 0 & n_i \leq N_i, Q_i + \lambda_i(t)\tau \leq n_i\mu\tau \\ & \text{or, } n_i > N_i, Q_i + \lambda_i(t)\tau \leq N_i\mu\theta + n_i\mu(\tau - \theta) \\ Q_i + \lambda_i(t)\tau - n_i\mu\tau & n_i \leq N_i, Q_i + \lambda_i(t)\tau > n_i\mu\tau \\ Q_i + \lambda_i(t)\tau - (N_i\mu\theta + n_i\mu(\tau - \theta)) & n_i > N_i, Q_i + \lambda_i(t)\tau > N_i\mu\theta + n_i\mu(\tau - \theta), \end{cases} \tag{5.4}$$

$$w_i = \begin{cases} 0 & n_i \leq N_i, Q_i = 0, \lambda_i(t) \leq n_i\mu \\ & \text{or, } n_i > N_i, Q_i = 0, \lambda_i(t) \leq N_i\mu \\ \frac{Q_i^2}{2(n_i\mu - \lambda_i(t))} & n_i \leq N_i, Q_i > 0, Q_i + \lambda_i(t)\tau \leq n_i\mu\tau \\ \frac{Q_i^2}{2(N_i\mu - \lambda_i(t))} & n_i > N_i, Q_i + \lambda_i(t)\theta \leq N_i\mu\theta \\ \frac{2Q_i + \theta(\lambda_i(t) - N_i\mu)}{2}\theta + \frac{(Q_i + \theta(\lambda_i(t) - N_i\mu))^2}{2(n_i\mu - \lambda_i(t))} & n_i > N_i, Q_i + \lambda_i(t)\theta > N_i\mu\theta, \text{ and} \\ & Q_i + \lambda_i(t)\tau \leq N_i\mu\theta + n_i\mu(\tau - \theta) \\ \frac{Q_i + Q'_i}{2}\tau & n_i \leq N_i, Q_i + \lambda_i(t)\tau > n_i\mu\tau \\ \frac{2Q_i + \theta(\lambda_i(t) - N_i\mu)}{2}\theta + \frac{[Q_i + \theta(\lambda_i(t) - N_i\mu)] + Q'_i}{2}(\tau - \theta) & n_i > N_i, \\ & Q_i + \lambda_i(t)\tau > N_i\mu\theta + n_i\mu(\tau - \theta). \end{cases} \tag{5.5}$$

A reasonable question to raise is whether an allocation that minimizes the wait incurred over only the current decision epoch (a myopic, or "greedy" solution) might be optimal over the entire time-horizon. If this were the case, then a dynamic programming framework would not be needed, and the allocations could be selected on the basis of the current period alone without considering the future. However, a counterexample shows that this is not the case.

Consider two fluid queues, A and B, of lengths $Q_A = 75$ and $Q_B = 15$ at time t for which, for simplicity, there will be no future arrivals through the end of the day at time $t + 90$. Let μ, the service rate, be equal to 0.5 units per minute, decisions occur every 30 minutes, switches require 15 minutes to be completed, and suppose that there are $N = 2$ servers and both are allocated to Queue B immediately prior to time t. Our state S at time t is thus $(75, 15, (0, 2))$. Consider a greedy allocation that minimizes only the wait to be incurred over the immediate period, $(t, t + 30)$. There are three possible allocations to consider: (0,2), in which all of the servers remain at Queue B; (1,1), in which one server is switched from Queue B to Queue A; and (2,0), in which both servers leave Queue B to serve Queue A.

- **Allocation (0,2):** All 75 customers in Queue A will continue to wait in line from time t to $t + 30$, contributing a total wait of $(75)(30) = 2250$ person-minutes. Meanwhile, the two servers at Queue B, operating at rate $2\mu = 1$ person per minute, will deplete the queue by time $t + 15$, causing a waiting time of $\left(\frac{15}{2}\right)(15) = 112.5$ person-minutes. **The total waiting time in this case is 2362.5 person-minutes.**

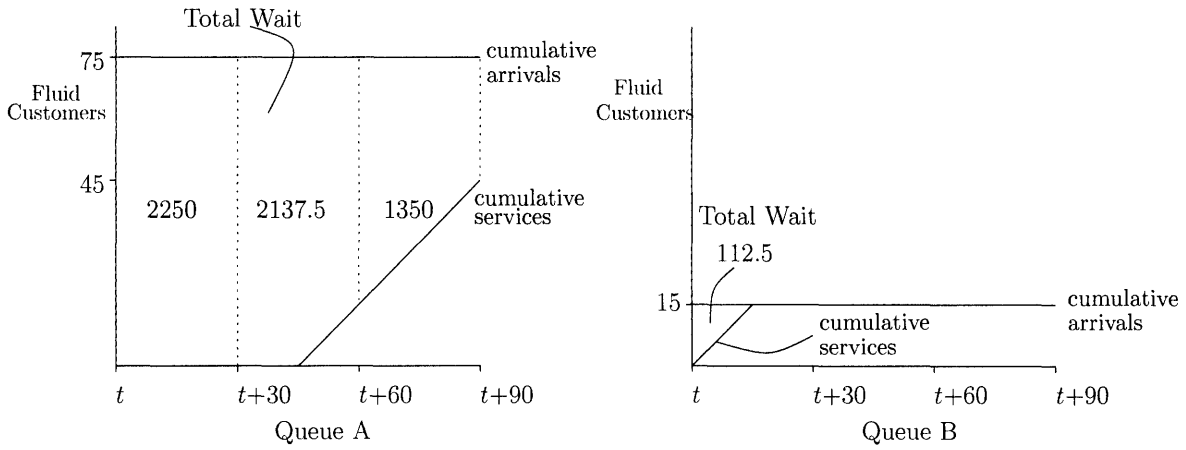

Figure 5-5: System evolution under greedy server allocations: (0,2), (2,0), (2,0). Although the waiting time during the period $(t, t+30)$ is minimized, the total wait over $(t, t+90)$ is not, as demonstrated by the optimal allocation shown in Figure 5-6.

- **Allocation (1,1):** From time t to $t+15$, the customers in Queue A will wait unserved before the switched server arrives. Over this period of time, the wait will be (75)(15) = 1125 person-minutes. From time $t+15$ to $t+30$, they will be served at rate 0.5 units per minute (7.5 units total), and the queue length at the end of the period will be 67.5. The waiting time in Queue A during this half of the period will therefore be $\left(\frac{75+67.5}{2}\right)(15) = 1068.75$ person-minutes. In Queue B, the switched server will be lost immediately, and the 15 customers in Queue B will be served at rate 0.5 units per minute until time $t+30$, at which point the queue will be depleted. The wait for Queue B will be $\left(\frac{15}{2}\right)(30)$=225 person-minutes. **The total waiting time in this case is 2418.75 person-minutes.**

- **Allocation (2,0):** If both servers are switched from Queue B to Queue A, then the customers in Queue A must wait 15 minutes before the switched servers arrive on duty, incurring a total wait of (75)(15)=1125 person-minutes. But once the servers arrive on duty, they will operate at a combined rate of 1 unit per minute, and will serve 15 of the 75 customers. The waiting time for this half of the period will thus be $\left(\frac{75+60}{2}\right)(15) = 1012.5$ person-minutes. Meanwhile, Queue B will have no servers on duty, and the 15 customers there will incur a wait of (15)(30)=450 person-minutes. **The total waiting time of this allocation is 2587.5 person-minutes.**

Because Allocation (0,2) yields the smallest total wait over the period $(t, t+30)$, the best greedy choice is to maintain the initial allocation of zero servers at Queue A and two servers at Queue B. Following this allocation, the system state S at time $t+30$ will be $(75, 0, (0, 2))$. Simple calculations show that the greedy allocation at both times $t+30$ and $t+60$ is (2,0), for a total wait over the period $(t, t+90)$ of 5850 person-minutes. Figure 5-5 shows the evolution of both queues as a result of this greedy allocation.

However, solving the dynamic program reveals that the optimal allocation is to select

	Greedy			Optimal		
Time	Initial State	Allocation Chosen	Wait Incurred	Initial State	Allocation Chosen	Wait Incurred
t	(75,15,(0,2))	(0,2)	2362.5	(75,15,(0,2))	(1,1)	2418.75
$t+30$	(75,0,(0,2))	(2,0)	2137.5	(67.5,0,(1,1))	(2,0)	1743.75
$t+60$	(60,0,(2,0))	(2,0)	1350	(45,0,(2,0))	(2,0)	900
TOTAL:			5850			5062.5

Table 5.2: System evolution and waiting time incurred under a Greedy Allocation (minimizing only the current period's waiting time) and an Optimal Allocation (minimizing the waiting time incurred over the entire interval from time t to the end of the time horizon at time $t+90$). Although the Greedy Allocation yields a smaller waiting time in the first period, the Optimal Allocation yields a smaller overall wait.

(1,1) at time t rather than (0,2), and then to select allocation (2,0) at times $t+30$ and $t+60$, yielding a total wait of only 5062.5 person-minutes! The evolution of the queues under the optimal allocation is shown in Figure 5-6, and the optimal allocation is compared with the greedy-allocation in Table 5.2. We can see that although the greedy method selects an allocation at time t that yields a smaller waiting time over the period $(t, t+30)$ than does the optimal allocation, by choosing instead allocation (1,1) in the first period, the system would experience lower waits in the long run.

Thus, it is insufficient to determine server allocations based on the current period alone, and the aforementioned dynamic program must be solved to find an optimal schedule allocation. Throughout this chapter, we will refer to this schedule allocation as the *original allocation*. Because our objective is to evaluate the improvement offered by reallocating servers dynamically as opposed to adhering to a predetermined schedule, we will use this original schedule allocation to test stochastic variants of the model, comparing its performance under stochasticity to that of their own best dynamic allocations.

5.3.2 Deterministic disruptions to passenger entry pattern

Though an airport might have estimates of half-hourly and hourly passenger entry rates, there can be disruptions (such as forecasted storms) that might cause these estimates to be higher or lower than expected. We would like to know whether changing the server allocation in response to such disruptions could be beneficial.

The first model we consider examines such disruptions on a deterministic level: suppose that at the start of the day it is expected that there will be bad weather throughout the day. During a bad weather scenario, we assume that a fraction of passengers might arrive at the airport an hour or two early to try to get on an earlier flight and avoid missed connections caused by delays. Still others might show up somewhat later than expected because of traffic delays. How does such a disruption alter the allocations and expected waiting time?

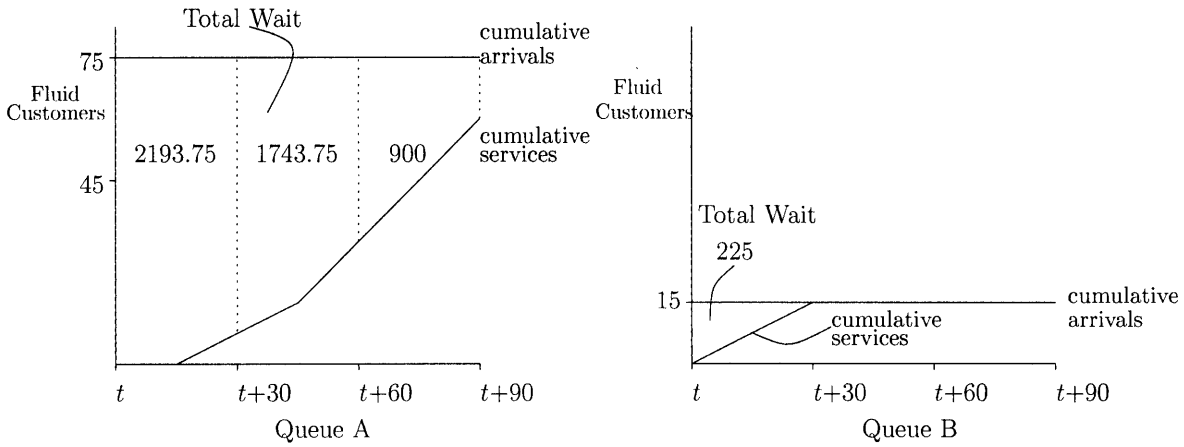

Figure 5-6: System evolution under optimal server allocations: (1,1), (2,0), (2,0). The waiting time in the period $(t, t+30)$ is not as small as under the Greedy allocation, but because of this initially suboptimal allocation, the total wait over $(t, t+90)$ is minimized.

We assume that during bad weather, 65% of the expected passengers enter the queue in their expected time period, while 5% enter two hours early, 10% an hour and a half early, 10% an hour early, 5% a half hour early and 5% a half hour late, with adjustments for boundary conditions at the start and end of the day. If decision epochs occur every 30 minutes ($\tau = 30$), then we have for the revised entry process $\hat{\lambda}_i(t)$:

$$\begin{aligned}
\hat{\lambda}_i(0) &= 0.95\lambda_i(0) + 0.05\lambda_i(30) + 0.10\lambda_i(60) + 0.10\lambda_i(90) + 0.05\lambda_i(120) \\
\hat{\lambda}_i(30) &= 0.05\lambda_i(0) + 0.90\lambda_i(30) + 0.05\lambda_i(60) + 0.10\lambda_i(90) + 0.10\lambda_i(120) + 0.05\lambda_i(150) \\
\hat{\lambda}_i(60) &= 0.05\lambda_i(30) + 0.80\lambda_i(60) + 0.05\lambda_i(90) + 0.10\lambda_i(120) + 0.10\lambda_i(150) + 0.05\lambda_i(180) \\
\hat{\lambda}_i(90) &= 0.05\lambda_i(60) + 0.70\lambda_i(90) + 0.05\lambda_i(120) + 0.10\lambda_i(150) + 0.10\lambda_i(180) + 0.05\lambda_i(210) \\
\hat{\lambda}_i(t) &= 0.05\lambda_i(t-30) + 0.65\lambda_i(t) + 0.05\lambda_i(t+30) + 0.10\lambda_i(t+60) + 0.10\lambda_i(t+90) \\
&\quad +0.05\lambda_i(t+120), \forall t = 120, 150, ..., T-150 \\
\hat{\lambda}_i(T-120) &= 0.05\lambda_i(T-150) + 0.65\lambda_i(T-120) + 0.05\lambda_i(T-90) + 0.10\lambda_i(T-60) \\
&\quad +0.10\lambda_i(T-30) \\
\hat{\lambda}_i(T-90) &= 0.05\lambda_i(T-120) + 0.65\lambda_i(T-90) + 0.05\lambda_i(T-60) + 0.10\lambda_i(T-30) \\
\hat{\lambda}_i(T-60) &= 0.05\lambda_i(T-90) + 0.65\lambda_i(T-60) + 0.05\lambda_i(T-30) \\
\hat{\lambda}_i(T-30) &= 0.05\lambda_i(T-60) + 0.70\lambda_i(T-30),
\end{aligned} \tag{5.6}$$

where $\lambda_i(s)$ refers to the passenger entry rate over the period $(s, s+30)$.

Using again a deterministic fluid framework on this problem with modified customer entry rates, we evaluate the performance of the best *fixed allocation*, the *original schedule allocation* from the previous section applied to this new entry pattern, and a *new schedule allocation* obtained by solving the dynamic program of (5.3) using entry rates adjusted for this bad weather scenario and true values in place of expectation.

5.3.3 Stochastic disruptions to passenger entry pattern

However, most weather disruptions (or other changes to customer entry rates) are not deterministic. Suppose now that a prediction of bad weather occurs during a period with probability p and that once it occurs, the customer entry rates over only the next four hours are affected in a manner similar to that described above. Once the bad weather period has passed, we assume that there will not be another bad weather period that day. For a bad weather watch beginning at time $s \geq 0$ well before the end of the day, we have:

$$\begin{aligned}
\hat{\lambda}_i(t) &= \lambda_i(t), \forall 0 \leq t < s, \forall s + 240 \leq t \leq T - 30 \\
\hat{\lambda}_i(s) &= 0.95\lambda_i(s) + 0.05\lambda_i(s+30) + 0.10\lambda_i(s+60) + 0.10\lambda_i(s+90) + 0.05\lambda_i(s+120) \\
\hat{\lambda}_i(s+30) &= 0.05\lambda_i(s) + 0.9\lambda_i(s+30) + 0.05\lambda_i(s+60) + 0.10\lambda_i(s+90) + 0.10\lambda_i(s+120) \\
&\quad +0.05\lambda_i(s+150) \\
\hat{\lambda}_i(s+60) &= 0.05\lambda_i(s+30) + 0.8\lambda_i(s+60) + 0.05\lambda_i(s+90) + 0.10\lambda_i(s+120) \\
&\quad +0.10\lambda_i(s+150) + 0.05\lambda_i(s+180) \\
\hat{\lambda}_i(s+90) &= 0.05\lambda_i(s+60) + 0.7\lambda_i(s+90) + 0.05\lambda_i(s+120) + 0.10\lambda_i(s+150) \\
&\quad +0.10\lambda_i(s+180) + 0.05\lambda_i(s+210) \\
\hat{\lambda}_i(s+120) &= 0.05\lambda_i(s+90) + 0.65\lambda_i(s+120) + 0.05\lambda_i(s+150) + 0.10\lambda_i(s+180) \\
&\quad +0.10\lambda_i(s+210) \\
\hat{\lambda}_i(s+150) &= 0.05\lambda_i(s+120) + 0.65\lambda_i(s+150) + 0.05\lambda_i(s+180) + 0.10\lambda_i(s+210) \\
\hat{\lambda}_i(s+180) &= 0.05\lambda_i(s+150) + 0.65\lambda_i(s+180) + 0.05\lambda_i(s+210) \\
\hat{\lambda}_i(s+210) &= 0.05\lambda_i(s+180) + 0.7\lambda_i(s+210).
\end{aligned} \tag{5.7}$$

If the bad weather watch begins within four hours from the end of the day, we have the additional boundary conditions that $\lambda(t) = 0$, for $t \geq T$, and that the 5% of the last period's customers that would otherwise arrive late due to the weather must arrive during their originally scheduled period.

In this version of the problem, the decision of when to switch servers depends not only on queue lengths, the current server allocation and expected future customer entry rates, but also on whether or not bad weather has been announced and if so, at what time. This is now a stochastic dynamic program of the form:

$$\begin{aligned}
W_T(.) &= 0 \\
W_t(S, t_{bw}) &= \min_{(n_A, n_B)} E\left[w_t\left(S, (n_A, n_B), \theta, t_{bw}\right) + W_{t+\tau}\left(\hat{S}_{t+\tau}(S, (n_A, n_B), \theta, t_{bw}), \hat{t}_{bw}\right)\right],
\end{aligned} \tag{5.8}$$

where t_{bw} equals -1 if bad weather has not yet been announced, and equals s if bad weather was announced at time $s \geq 0$. $\hat{t}_{bw}$ is the new value of t_{bw} in the next decision period, depending on whether bad weather is announced during the current period (with probability

p) or not. The expectation in (5.8) is taken with respect to t_{bw}, yielding the following:

$$\begin{aligned}
W_T(.) &= 0 \\
W_t(S,-1) &= \min_{n_A,n_B} w_t\left(S,(n_A,n_B),\theta,-1\right) + pW_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,-1),t+\tau\right) \\
&\quad +(1-p)W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,-1),-1\right) \\
W_t(S,s) &= \min_{n_A,n_B} w_t\left(S,(n_A,n_B),\theta,s\right) + W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,s),s\right), s \geq 0.
\end{aligned} \tag{5.9}$$

Because the evolution of the system over time is no longer deterministic, the above DP yields a *dynamic allocation* that changes not only for different decision epochs but also dependent on the queue lengths and weather pattern observed. We can compare the performance of this dynamic allocation to that of the best *fixed allocation* and the *original schedule allocation* applied to this stochastic system.

5.3.4 Randomized entry rates

Another way to introduce stochasticity to the entry pattern is to consider entry rates that are, independently in each period and for each checkpoint, a factor $(1-\alpha)$ lower or $(1+\alpha)$ higher than originally estimated, for $0 \leq \alpha \leq 1$, with probability β (such that $\beta \leq 0.5$). Such a situation might arise if the estimated entry rates are not very accurate, or if gate changes or flight cancellations change the number of passengers passing through the security checkpoints. High values of α and β indicate greater uncertainty about the entry rates.

The dynamic program is given by:

$$\begin{aligned}
W_T(.) &= 0 \\
W_t(S) &= \min_{(n_A,n_B)} \beta^2 \left[w_t\left(S,(n_A,n_B),\theta,(1-\alpha)\lambda_{A,t},(1-\alpha)\lambda_{B,t}\right)\right. \\
&\quad +W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,(1-\alpha)\lambda_{A,t},(1-\alpha)\lambda_{B,t})\right) \\
&\quad +w_t\left(S,(n_A,n_B),\theta,(1-\alpha)\lambda_{A,t},(1+\alpha)\lambda_{B,t}\right) \\
&\quad +W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,(1-\alpha)\lambda_{A,t},(1+\alpha)\lambda_{B,t})\right) \\
&\quad +w_t\left(S,(n_A,n_B),\theta,(1+\alpha)\lambda_{A,t},(1-\alpha)\lambda_{B,t}\right) \\
&\quad +W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,(1+\alpha)\lambda_{A,t},(1-\alpha)\lambda_{B,t})\right) \\
&\quad +w_t\left(S,(n_A,n_B),\theta,(1+\alpha)\lambda_{A,t},(1+\alpha)\lambda_{B,t}\right) \\
&\quad \left.+ W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,(1+\alpha)\lambda_{A,t},(1+\alpha)\lambda_{B,t})\right)\right] \\
&\quad +\beta(1-2\beta)\left[w_t\left(S,(n_A,n_B),\theta,(1-\alpha)\lambda_{A,t},\lambda_{B,t}\right)\right. \\
&\quad +W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,(1-\alpha)\lambda_{A,t},\lambda_{B,t})\right) \\
&\quad \left.+w_t\left(S,(n_A,n_B),\theta,(1+\alpha)\lambda_{A,t},\lambda_{B,t}\right) + W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,(1+\alpha)\lambda_{A,t},\lambda_{B,t})\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&+w_t\left(S,(n_A,n_B),\theta,\lambda_{A,t},(1-\alpha)\lambda_{B,t}\right)+W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,\lambda_{A,t},(1-\alpha)\lambda_{B,t})\right)\\
&+w_t\left(S,(n_A,n_B),\theta,\lambda_{A,t},(1+\alpha)\lambda_{B,t}\right)+W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,\lambda_{A,t},(1+\alpha)\lambda_{B,t})\right)\Big]\\
&+(1-2\beta)^2\left[w_t\left(S,(n_A,n_B),\theta,\lambda_{A,t},\lambda_{B,t}\right)+W_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta,\lambda_{A,t},\lambda_{B,t})\right)\right],
\end{aligned}
\tag{5.10}
$$

where the functions $w_t(S)$ and $W_t(S)$ depend on the actual arrival rate experienced at each time and checkpoint. Again, we compare the waiting time incurred in the solution of the dynamic program to those under fixed and schedule allocations.

5.3.5 Stochastic service times

We now return to deterministic entry rates and consider instead stochasticity in the service process. We have been assuming that each unit of flow, representing a customer, has a constant service time equal to $1/\mu$. However, different customers can take different amounts of time to pass through security, and we can model this by assigning to each unit of fluid, k, in a queue i an exponentially distributed service time X_k with mean $1/\mu$ minutes, which represents the total amount of "work" that must be done on that unit of flow. However, that work is divided over all N_i servers on-duty at that queue, such that the unit of fluid is processed at a uniform rate equal to N_i/X_k, and when completed, a new service time for the next unit of flow is randomly generated. Thus, we maintain the continuous fluid form of the model while stochastically changing, on a unit-by-unit basis, the rate at which the fluid is processed.

We can use (5.3) to describe the optimal dynamic allocation for this case where service times are stochastic, taking the expectation over the service time process, and augmenting the system state S from the deterministic fluid model to include the remaining service time of the unit of fluid currently being processed. We can then compare the solution given by the DP to a best *fixed allocation* and to the *original schedule allocation* of the deterministic fluid model.

5.4 Approximate dynamic programming solution techniques

Although the fluid formulations allow us to ignore the individual processing of customers (except in the stochastic service times variant), we must still keep track, at each time point, of the amount of fluid in each queue and of the current allocation, causing the state space in (5.3) to become prohibitively large. Furthermore, the nonstationarity of the entry processes over the course of the day forces the wait-to-go function to be dependent on time, preventing steady-state conditions from setting in which would simplify the system of equations. As a result of this, at the k^{th} decision epoch, there could be as many as $(\min(N_{max_A},N) - \max(N-N_{max_B},0)+1)^{k-1}$ possible combinations of queue lengths and server allocations

to consider in the deterministic setting, and an infinite number of states under stochastic service times. If we consider decision epochs arising every half hour over the course of an entire day, this can get prohibitively large, even in the deterministic model, and it becomes even harder if decision epochs occur more frequently.

Thus, in order to be able to solve the models described in the previous sections, we must resort to approximate dynamic programming techniques. (A detailed discussion of such techniques is provided in chapter six of [25]). For all of the models considered (deterministic and stochastic), we rely upon state space aggregation and restricting the number of periods into the future explored by the dynamic program. For the stochastic service model, we also use simulation and reduce the solution space considered in order to facilitate computation.

5.4.1 Aggregated state space with limited lookahead

To reduce the number of different states per stage requiring evaluation, we aggregate queue lengths into multiples of ten. During a decision epoch, the current period wait incurred, w, is computed exactly, as described in (5.5), but to compute the wait-to-go starting from the next decision epoch, the queue lengths Q'_A and Q'_B at the end of the period are rounded to the nearest multiple of ten. Because the fluid model can yield fractional values of Q'_A and Q'_B, this rounding can significantly reduce the number of states visited by the DP. Furthermore, since queue lengths in an airport setting are generally not monitored precisely, choosing server allocations based on rough estimates of queue lengths is a realistic approximation.

An additional difficulty stems from dynamic programming's "curse of dimensionality": the number of possible solutions (and states to explore) grows exponentially with the time horizon considered. The optimal server allocation at time t depends on the repercussions such an allocation will cause through the remainder of the time horizon T. However, although a dynamic program takes into account the full extent of such repercussions, realistically it seems improbable that a decision made at 6:30AM could cause a significant impact on the state of the system at 4:00PM, particularly if reallocations can be made every half- or quarter-hour. So we also use a *limited lookahead* approximation, where rather than determining an allocation based on the total wait-to-go for the remainder of the time horizon, T, we consider only the next two or three hours of the time horizon (in most of our tests, a three hour lookahead is used with thirty minute decision epochs, while in one test having fifteen minute decision epochs, a two hour lookahead is used). If we let L be the length of the lookahead period, then at time t and state S, the allocation selected is the $\overline{(n_A, n_B)}_{S,t}$ that achieves the minimum in

$$\min_{(n_A,n_B)} E\left[w_t\left(S,(n_A,n_B),\theta\right) + \tilde{W}_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta)\right)\right], \tag{5.11}$$

where

$$\begin{aligned} \tilde{W}_{t+L}(.) &= 0 \\ \tilde{W}_t(S) &= \min_{(n_A,n_B)} E\left[w_t\left(S,(n_A,n_B),\theta\right) + \tilde{W}_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta)\right)\right]. \end{aligned} \tag{5.12}$$

Although the policy $\overline{(n_A, n_B)}_{S,t}$ is computed using a truncated wait-to-go function, $\tilde{W}_t(S)$, the true wait-to-go from any time t and state S, $W_t(S)$, is obtained by stepping through the system using the best allocations $\overline{(n_A, n_B)}_{S,t}$ found above:

$$\begin{aligned} W_T(.) &= 0 \\ W_t(S) &= E\left[w_t\left(S, \overline{(n_A, n_B)}_{S,t}, \theta\right) + W_{t+\tau}\left(\hat{S}(S, \overline{(n_A, n_B)}_{S,t}, \theta)\right)\right]. \end{aligned} \tag{5.13}$$

For the first two models, having deterministic entries and services, we can ignore the expectation in equations (5.11)-(5.13). For the model with stochastically arising disruptions to the entry process, we note that although the bad weather may arise unexpectedly, the entries and services are deterministic once the weather status is revealed. Thus, the only uncertainty is when, if ever, bad weather will start. Because it can start in any one of the decision epochs (or never start), there are only $\lceil T/\tau \rceil + 1$ stochastic trajectories to explore, rendering the expectations in equations (5.11)-(5.13) relatively simple to compute. So state aggregation with a limited lookahead period is sufficient for exploring the deterministic models and the model with stochastic disruptions to the entry process.

5.4.2 Approximating expected values

By contrast, calculating the expected value in the above equations is significantly more difficult in the case of stochastic service times. Because each unit of flow is assigned its own exponentially distributed service time, there are an infinite number of trajectories to consider. To address this, we implement two different heuristics, both involving simulation rather than exact computation of the expected value.

A hybrid allocation

The first of these is a "hybrid" of a stochastic dynamic program and a deterministic dynamic program, in which the allocation at each state and stage is made based on the deterministic version of the dynamic program in (5.11)-(5.12) (i.e. the decision is based on the assumption that the system evolves deterministically), but in which the actual evolution of the system (as in (5.13)) is stochastic, allowing the system to visit states not visited by the original deterministic DP and from those states to select allocations different from those in the original schedule allocation. (This is similar in spirit to the BIGSTEP method of [60] in which the system status is reviewed at a decision epoch, and based on this system state, a new processing plan is determined from a heuristic and is used for that period, over which the process evolves stochastically until the next decision point). Rather than computing the expected value in (5.13) exactly, we simulate the evolution of the system. For iterations $k = 1...K$ having stochastic trajectory ω_k, we compute $W_0(S_0)_k$ from an initial state S_0 recursively, as follows:

- Function $W_t(S)_k$:

    1. If $t = T$, return 0. Otherwise,

2. Find $\overline{(n_A, n_B)}_{S,t}$ which is

$$\arg \min_{(n_A,n_B)} w_t\left(S,(n_A,n_B),\theta\right) + \tilde{W}_{t+\tau}\left(\hat{S}_{t+\tau}(S,(n_A,n_B),\theta)\right),$$

where

$$\begin{aligned} \tilde{W}_{t+L}(.) &= 0 \\ \tilde{W}_t(S) &= \min_{(n_A,n_B)} w_t\left(S,(n_A,n_B),\theta\right) + \tilde{W}_{t+\tau}\left(\hat{S}_{t+\tau}\left(S,(n_A,n_B),\theta\right)\right). \end{aligned}$$

is the solution to the deterministic fluid flow problem with limited lookahead L.

3. Return

$$W_t(S)_k = w_t\left(S,\overline{(n_A,n_B)}_{S,t},\theta,\omega_k\right) + W_{t+\tau}\left(\hat{S}_{t+\tau}\left(S,\overline{(n_A,n_B)}_{S,t},\theta,\omega_k\right)\right)_k.$$

The average total waiting time is $W_0(S_0) = \frac{1}{K}\sum_k W_0(S_0)_k$. It is in Step 2 above that the allocation is selected based on the deterministic model, and it is in Step 3 that the system evolves stochastically, according to ω_k, resulting in a new state $\hat{S}$ from which to recurse.

Nearest neighbor solutions

Although the above hybrid allocation introduces some stochasticity in the form of system evolution between allocations, ideally we would like to allow the allocations chosen by the DP in Step 2 above to consider the stochastic evolution. We again use simulation to achieve this, by selecting in Step 2, for each state encountered, an allocation that minimizes the average limited lookahead wait-to-go over several simulation runs. Once an allocation has been chosen, we simulate the system evolution until the next decision epoch, as we did in Step 3 of the hybrid heuristic. Note that by simulating the lookahead period, we are performing several times the very step that takes the longest to evaluate. In this heuristic, therefore, we choose to restrict the solution space to a subset $(n_A, n_B)'$ of "reasonable" allocations by recognizing that the optimal stochastic allocation is likely to be close to the original deterministic schedule allocation. As such, if the original schedule called for an allocation of (n_A, n_B), then our allocation in the stochastic model must be a "nearest neighbor" of (n_A, n_B): $(n_A, n_B)' = \{(n_A - 1, n_B + 1), (n_A, n_B), (n_A + 1, n_B - 1)\}$. (Using a neighborhood of ± 2 servers did not appreciably change the results on our data. The total number of servers in our data is relatively small so that each additional server shifted caused a disproportionate shift in load balance).

The heuristic is as follows. For general iterations $k = 1...K$ having stochastic trajectory ω_k, and lookahead iterations $j = 1...J$ within each general iteration k having stochastic trajectory ψ_{kj} we compute $W_0(S_0)_k$ from an initial state S_0 recursively, as follows:

- Function $W_t(S)_k$:

  1. If $t = T$, return 0. Otherwise,

2. Find $\overline{(n_A, n_B)}_{S,t}$ which is

$$\arg \min_{(n_A,n_B)\in(n_A,n_B)'} \quad \frac{1}{J}\sum_j \left(w_t\left(S,\left(n_A,n_B\right),\theta,\psi_{kj}\right)\right.$$
$$\left. + \tilde{W}_{t+\tau,j,k}\left(\hat{S}_{t+\tau}\left(S,\left(n_A,n_B\right),\theta,\psi_{kj}\right)\right)\right),$$

where

$$\begin{aligned}\tilde{W}_{t+L,.,.}(.) &= 0 \\ \tilde{W}_{t,j,k}(S) &= \min_{(n_A,n_B)\in(n_A,n_B)'} w_t\left(S,\left(n_A,n_B\right),\theta,\psi_{kj}\right) \\ &\quad +\tilde{W}_{t+\tau,j,k}\left(\hat{S}_{t+\tau}\left(S,\left(n_A,n_B\right),\theta,\psi_{kj}\right)\right).\end{aligned}$$

3. Return

$$W_t(S)_k = w_t\left(S,\overline{(n_A,n_B)}_{S,t},\theta,\omega_k\right) + W_{t+\tau}\left(\hat{S}_{t+\tau}(S,\overline{(n_A,n_B)}_{S,t},\theta,\omega_k\right)_k.$$

Once again, the average total waiting time is $W_0(S_0) = \frac{1}{K}\sum_k W_0(S_0)_k$.

5.5 Data

To test these models, we use data provided by Massport, the Port Authority for Massachusetts that operates Boston Logan International Airport. This data contains throughput, estimated average wait and total number of lanes open at each security checkpoint at Logan airport, on an hourly or half-hourly basis, for the week of January 18-24, 2005. Unless otherwise stated, we use the January 18 data.

Logan Airport is divided into five terminals, Terminal A (which recently opened in March 2005 and serves Delta Air Lines), Terminals B and C which handle primarily domestic flights, Terminal D which handles domestic flights primarily for Air Tran Airways and Terminal E which is the international terminal but which also accommodates a few domestic departures (See Figure 5-7). Each terminal has at least one security checkpoint, each serving subsets of gates within the terminal: Terminal A has one, Terminal B has seven (although only six were ever open at the time the data provided to us were collected), Terminal C has three, Terminal D one, and Terminal E two, and each checkpoint has a certain number of lanes through which passengers pass (ranging from one lane at Terminal D's checkpoint to eight lanes at Terminal A's checkpoint). The checkpoints are equipped with an x-ray bag scanner for screening carry-on luggage and a metal detector gate through which passengers walk. We use the term *server* to refer to an open lane including equipment (x-ray machine and metal detector) and manpower (employees to operate the equipment, direct passengers through the lane and conduct searches). For example, at checkpoint E2, while there is capacity for up to seven open lanes, at any given time, there might be only five "servers" on-duty,

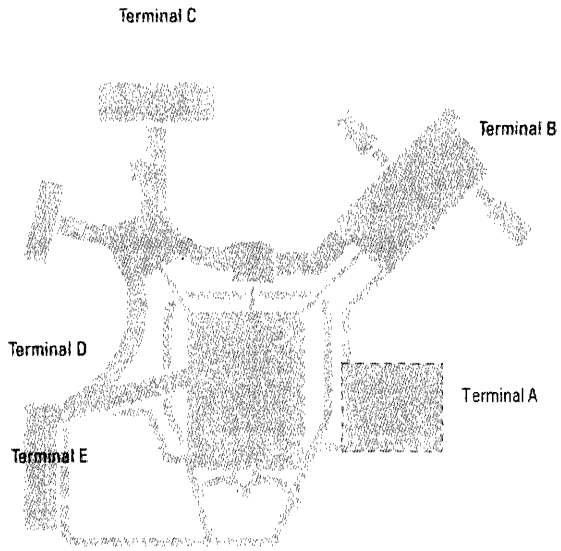

Figure 5-7: Diagram of Boston's Logan International Airport. *Source: Delta Air Lines website (edited)*, http://www.delta.com/travel/before/airport_info/airport_maps/boston_bos/index.jsp

corresponding to five sets of x-ray machines, metal detectors and screening employees.

Unless explicitly stated, our analysis uses the data of Terminals C and E, treating Terminal C's three checkpoints (having a total of eleven possible screening lanes) as one mega-checkpoint, and treating Terminal E's two checkpoints (having a total of nine lanes) as one. Any additional data from other sets of terminals or other days used for sensitivity analysis will be specifically indicated.

We estimate the entry rate, $\lambda_i(t)$ for any terminal i from the throughput information provided by Massport. We let $\lambda_i(t)$ be equal to the average throughput per minute recorded during the corresponding half-hour block, and if the data at a particular time are given for an hour-long block, we divide the stated throughput by two to estimate the half-hourly rate. These estimated entry rates for Terminals C and E are shown in Table 5.3 for the period from 5:00 AM to 6:30 PM.

The service rate, μ, is harder to estimate. According to Massport, the maximum possible service rate, μ, of a single x-ray machine and metal detector is approximately two hundred passengers per hour (3.3 passengers per minute). However, in the data, the stated throughput (TP) is rarely ever this high. While $\frac{TP}{n}$ estimates the service rate of each of n servers on-duty during the corresponding half-hour period, this does not indicate the service *capacity* because at some times there might have been one or more servers idle. One remedy is to consider only those periods where the average waiting time stated in the data is nonzero, as the formation of queues indicates servers were likely working near maximum capacity. During these periods, the average throughput is closer to 2.8 passengers per minute. Although this estimate varies greatly across time of day and terminals (for instance, if any of the machines were set to be more sensitive than usual to triggers, or if screeners were encouraged to work

Time	Terminal C	Terminal E	Time	Terminal C	Terminal E
5:00 AM	6.20	2.90	12:00 PM	8.70	1.98
5:30	7.23	7.2	12:30	11.68	4.05
6:00	16.97	6.93	1:00	11.68	4.05
6:30	14.70	8.73	1:30	17.13	2.85
7:00	21.00	8.13	2:00	17.13	2.85
7:30	26.77	7.23	2:30	14.35	7.45
8:00	25.30	7.23	3:00	14.35	7.45
8:30	17.03	4.50	3:30	19.03	9.17
9:00	13.27	5.43	4:00	22.47	9.97
9:30	12.52	4.28	4:30	15.83	10.60
10:00	12.52	4.28	5:00	13.93	11.47
10:30	12.38	4.97	5:30	8.80	13.00
11:00	12.38	4.97	6:00	10.57	10.50
11:30	8.70	1.98			

Table 5.3: Customer entry rates (in customers per minute) to Terminals C and E at Boston Logan International Airport on January 18, 2005, from 5:00 AM to 6:30 PM

more quickly during peak hours to accommodate the additional load, service rates might be lower or higher than usual), it seems to be a reasonable estimate of the average service rate, based on calibration simulations[1].

We consider decision epochs of 30 minutes in general ($\tau = 30$), but we also explore whether more frequent epochs (e.g., $\tau = 15$) yield significant improvements. We consider a range of switching times, θ, equal to 0, 5, 10, 15, or 30 minutes, provided $\theta \leq \tau$.

Because this model holds the total number, N, of servers on-duty constant, we hypothesize that dynamic server allocation might be useful primarily on systems where the individual customer entry rates at the two queues vary in opposite senses to each other, but where the total number of servers needed to serve the two queues is relatively constant. According to the data from Logan airport, from 5:00 AM to 6:30 PM on January 18, 2005, the actual number of checkpoint lanes open at Terminals C and E at Logan airport stayed fairly constant around 10, though the entry rates at the two terminals varied, so we use $N = 10$ in our trials.

In summary, we examine two queues representing Terminals C and E at Logan Airport

[1]Airports such as Logan may occasionally use different checkpoint configurations to accommodate larger volumes of passengers. Most of the time, each lane has one metal detector and one x-ray bag scanner. Those familiar with United States airports will agree that walking through the metal detector is generally a rapid process, while waiting for one's bags, coats, laptops and shoes to pass through the x-ray machine can take much longer, causing it to be the bottleneck stage in the screening process. To compensate during periods of heavy traffic, Logan occasionally uses a 2-1 configuration, meaning they use two x-ray machines in parallel with a single metal detector. Successive customers use alternate x-ray machines for their bags but pass through the same metal detector. Though the use of such configurations can increase the service rate, μ, we do not consider such configurations in our model.

		$\tau = 30$			$\tau = 15$		
θ	Fixed Alloc.	Schedule Alloc.	Number of Switches	% Imprvmt.	Schedule Alloc.	Number of Switches	% Imprvmt.
0	4.135 (7,3)	1.886	18	54.38	1.846	16	55.35
5		2.089	8	49.47	2.061	10	50.15
10		2.298	8	44.44	2.257	11	45.41
15		2.486	9	39.88	2.506	11	39.39
30		3.051	9	26.21	N/A	N/A	N/A

Table 5.4: Average waiting times (minutes per customer) under the best fixed allocation, $\overline{(n_A, n_B)} = (7, 3)$, and schedule allocations with decision epochs every thirty minutes and every fifteen minutes; the total number of servers switched and the percentage improvement of the schedule allocations over the fixed allocation, by switching time, θ.

that share $N = 10$ servers who are each capable of processing 2.8 units of flow (customers) per minute and where flow enters the queue at the rates given in Table 5.3. The servers can be reallocated every $\tau = 30$ minutes (unless otherwise indicated) and they require $\theta = 0, 5, 10, 15$ or 30 minutes to arrive to their new post after a reallocation. Using this data, we evaluate fixed, schedule and dynamic allocations for the cases of deterministic entries and services, deterministically disrupted entries, stochastically disrupted entries, and stochastic service times. The results are presented in the following section.

5.6 Results

5.6.1 Deterministic entries and services

Table 5.4 shows the performance, measured by the average waiting time per customer, of the different allocations in the simple fluid model with deterministic entries and services. The average waiting time under the best fixed allocation is roughly four minutes per customer, and is achieved by allocating seven servers to Terminal C and three servers to Terminal E (note that this average wait does not vary with the switching time θ because servers are not allowed to be switched under a fixed allocation). Nearly 70% of the total entries to the two queues, as given in Table 5.3, are to Terminal C, so a (7,3) fixed allocation is not surprising.

In a schedule allocation, where switches are allowed to occur every $\tau = 30$ minutes, the average waiting time decreases significantly as compared to the fixed allocation, ranging from just under 2 minutes per customer if switches are instantaneous to 2.5 minutes if they take 15 minutes. Even in the extreme case of 30 minute switching times, the waiting time per customer is still only approximately 3 minutes, meaning that even if switches take a very long time to complete, we would still prefer to switch than maintain a fixed allocation.

As we would expect, the number of switches is generally smaller under non-zero switching

Time	θ=0	θ=5	Time	θ=0	θ=5
5:00 AM	(3,7)	(3,7)	12:00 PM	(5,5)	(7,3)
5:30	(3,7)	(7,3)	12:30	(5,5)	(7,3)
6:00	(7,3)	(7,3)	1:00	(5,5)	(7,3)
6:30	(6,4)	(7,3)	1:30	(7,3)	(7,3)
7:00	(7,3)	(7,3)	2:00	(7,3)	(7,3)
7:30	(9,1)	(8,2)	2:30	(7,3)	(7,3)
8:00	(9,1)	(8,2)	3:00	(7,3)	(7,3)
8:30	(7,3)	(8,2)	3:30	(7,3)	(7,3)
9:00	(5,5)	(7,3)	4:00	(8,2)	(7,3)
9:30	(5,5)	(7,3)	4:30	(5,5)	(6,4)
10:00	(5,5)	(7,3)	5:00	(5,5)	(5,5)
10:30	(5,5)	(7,3)	5:30	(5,5)	(5,5)
11:00	(5,5)	(7,3)	6:00	(5,5)	(5,5)
11:30	(5,5)	(7,3)			

Table 5.5: Schedule allocations under the deterministic fluid model for $\tau = 30$ minute decision epochs and decision times of $\theta = 0$ or 5 minutes.

times than when switches can occur instantaneously. However, the magnitude of the switching time does not seem to influence the number of switches made as much as the existence of a switching delay. There is a sharp decrease in the number of switches required as θ increases from 0 to 5 minutes, but little change as the switching time increases from 5 to 30 minutes. Table 5.5 shows the allocations selected at each period for $\theta = 0$ and $\theta = 5$. First we note that the allocation when switches are instantaneous fluctuates a lot, particularly in the early morning peak period. By contrast, the allocation when $\theta = 5$ minutes remains relatively steady throughout the day. This suggests that when there is essentially no cost to switching, the model can accommodate minor imbalances between the two queues. However, if time is lost while switching, the model prefers to maintain service capacity by switching only when necessary. We note that in the early morning and the late afternoon, the two allocations are generally within ± 1 server of each other. From approximately 9 AM to 1:30 PM, however, the allocations differ by two servers. This should not be interpreted as a difference caused by the switching times. In fact, during this period, the total entry rate to both queues is significantly lower than the total service capacity available, so under either allocation ((5,5) or (7,3)), the waiting time incurred over this period is zero.

A somewhat surprising result is the *increase* in the number of switches that take place as θ increases from 10 minutes (in which the schedule switches 8 servers over the course of the day) to 15 minutes (in which the allocation has 9 switches). In fact, there are instances when switches *must* occur in order to accommodate sharp changes in the relative entry rates between the two queues. However, higher switching times cause there to be a longer delay before the switched servers become available. This extra delay can cause sufficiently longer queues to form at the recipient queue that the system must switch even more servers to this

			$\theta = 10$			$\theta = 15$		
Policy	**Time**	**Entry Rates**	**Initial Queue**	**End Queue**	**Wait**	**Initial Queue**	**End Queue**	**Wait**
(4,6)	5:30 PM	(8.8,13)	(55,8)	(0,0)	889.15	(55,18)	(0,0)	1369.16
(4,6)	6:00 PM	(10.6,10.5)	(0,0)	(0,0)	0	(0,0)	(0,0)	0
Total					889.15			**Opt. 1369.16**
(5,5)	5:30 PM	(8.8,13)	(55,8)	(0,6)	780.87	(55,18)	(0,30)	1325.87
(5,5)	6:00 PM	(10.6,10.5)	(0,6)	(0,0)	5.14	(0,30)	(0,0)	128.57
Total					**Opt. 786.01**			1454.44

Table 5.6: Total waiting times (person-minutes) incurred under two different allocations, (4,6) and (5,5), and two different switching times, $\theta = 10$ and $\theta = 15$ minutes, for the last two periods of the day when all previous allocations have been identical and the initial allocation at 5:30 PM is (6,4). The optimal allocation when $\theta = 10$ is to switch *one* server from A to B, while when $\theta = 15$, it is optimal to switch *two* servers from A to B.

queue to compensate. To see this, we discuss the cases of $\theta = 10$ and $\theta = 15$ in greater detail.

The two allocations are identical throughout the day until 5:30 PM. At this time, the initial allocation is (6,4), and when $\theta = 10$ the system has 55 customers at Queue A, 8 customers at Queue B, and for $\theta = 15$, Queue A has 55 customers, but Queue B has 18 customers (although the allocations up to this point are identical, different switching times cause the queue lengths to differ). The optimal allocation for the decision epoch 5:30-6:00 PM under $\theta = 10$ is to switch one server from Queue A to Queue B for a final allocation of (5,5), whereas for $\theta = 15$ the decision is to switch two servers from Queue A to Queue B for a final allocation of (4,6). This additional switch is due to the queue that develops during the extra time required by the servers to complete the switch[2]. Table 5.6 shows the system evolution for the two cases under the (4,6) and (5,5) allocations.

With respect to the timing of switches, we would generally expect the fraction of the N servers allocated to a queue i to be roughly proportional to the fraction of that period's total entry rate attributed to queue i. If we let $(\hat{N}_A, \hat{N}_B)$ be the number of servers we would expect at queues A and B based on the entry rate proportion, then we have:

$$(\hat{N}_A(t), \hat{N}_B(t)) = \left(\frac{\lambda_A(t)}{\lambda_A(t) + \lambda_B(t)} N, \frac{\lambda_B(t)}{\lambda_A(t) + \lambda_B(t)} N \right), \tag{5.14}$$

[2]One might suggest that the extra switch when $\theta = 15$ is due to the initial queue length at Queue B being larger when $\theta = 15$ than when $\theta = 10$. However, even if the initial queue lengths at 5:30 PM at Queue B for the two cases were both equal to 8, we would still find a (5,5) allocation to be optimal when $\theta = 10$ and a (4,6) allocation to be optimal when $\theta = 15$.

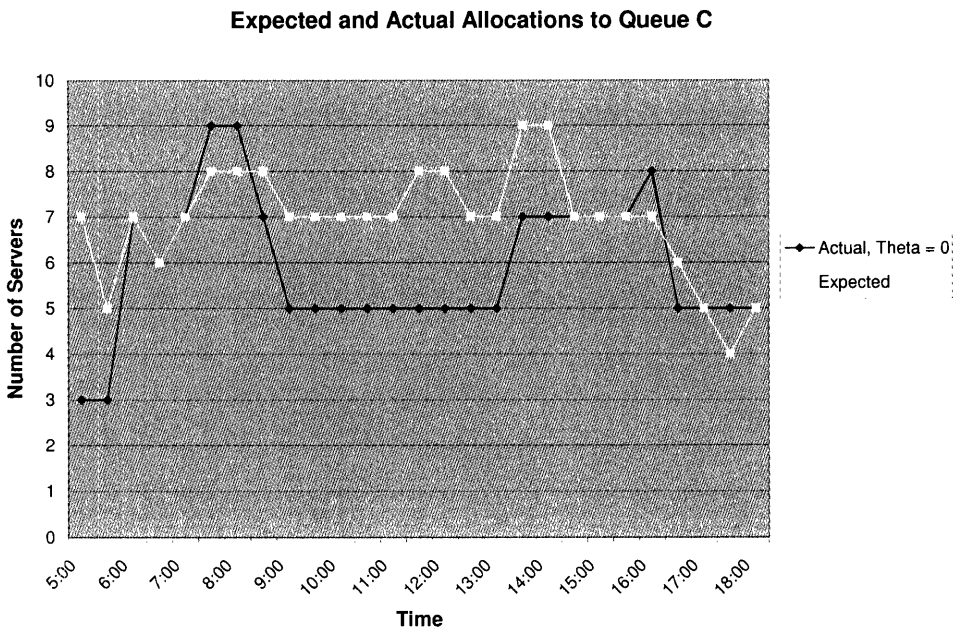

Figure 5-8: Expected and actual server allocations to Terminal C under the deterministic fluid model with zero switching times. There does not appear to be a strong correlation between the proportionate entry rate to Terminal C and the allocation chosen by the model.

where the values can be rounded to the nearest integers. However, if we plot $\hat{N}_A(t)$ for the Terminal C and E data (a mirror plot would be obtained for $\hat{N}_B(t)$), and compare the curve to the actual allocation selected by the program for the case of zero switching times (Figure 5-8), we see that the two curves are not very similar. First, we note that from 9:00 AM to 1:30 PM, the entry rates to both queues are lower than the service capacity provided by the ten servers considered, so during this period, several different allocations can be used and still maintain zero waiting times. However, even outside this period there does not appear to be a strong correlation between the actual allocations yielded by the model and those expected based on entry rates. There is, however, a closer correspondence of the proportionate entry rates to the allocations chosen by the models having *non-zero* switching times, as shown in Figure 5-9. It is difficult to ascertain whether switches occur earlier or later than in the expected allocation as a result of increased switching times. One might hypothesize that switches should occur earlier when switching times are high so that the switched servers can arrive at their new post before a surge in demand. While this does happen in some of the switches shown in Figure 5-9, there are others where a switch occurs later than the expected switch. An intuitive rule for timing switches based on θ is difficult to obtain.

Lastly, we see in Table 5.4 that there is not an appreciable difference in the waiting times achieved by selecting an allocation every thirty minutes versus every fifteen minutes in this deterministic case. Although, as expected, the fifteen minute allocation generally performs

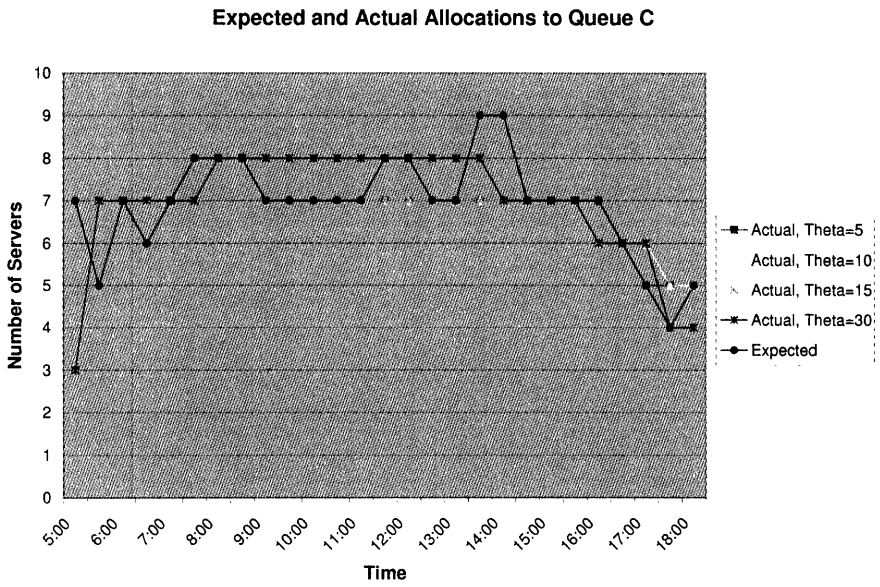

Figure 5-9: Expected and actual server allocations to Terminal C under the deterministic fluid model with non-zero switching times. Although the actual server allocations follow roughly the expected allocation based on proportionate entry rates, there does not appear to be a consistent effect of switching time on the timing of such switches.

somewhat better than the thirty-minute allocation, the approximate nature of the dynamic program used resulted in it faring worse for the case of $\theta = 15$. We also see that there is not a large difference between the number of switches taking place under a $\tau = 30$ minute decision epoch versus a $\tau = 15$ minute decision epoch, even though the latter case offers twice as many opportunities to switch. It seems that choosing a new allocation every fifteen minutes offers some moderate flexibility as to the timing of switches, but does not cause additional switches to take place, nor is it pivotal in reducing average waits. One possible explanation is that the passenger entry rates in the data are given in half-hour blocks. Because the entry rates do not change over the course of a thirty-minute decision epoch, offering more opportunities to switch during that period of time is perhaps unnecessary. It is possible that for quarter-hourly entry rates, for instance, having more frequent decision epochs might prove more useful. For the remainder of the analysis, we focus on $\tau = 30$ minute decision epochs.

5.6.2 Deterministically disrupted entries

In the case where the entry rates shift in a predictable pattern from the expected rates (for instance, due to a prediction of bad weather), we examine how the original schedule allocation from the deterministic "fair-weather" system compares to a new schedule allocation that is based on the expected changes. Table 5.7 shows the average waiting time under a fixed allocation, the original schedule allocation from the previous section and a new schedule

θ	Best Fixed Allocation	Original Schedule ($\tau = 30$)	New Schedule ($\tau = 30$)	Number of Switches	% Imprvmt. (fixed)	% Imprvmt. (original)
0	2.428 (7,3)	3.800	0.566	10	76.69	85.10
5		1.104	0.782	5	67.80	29.17
10		1.105	0.851	5	64.93	22.99
15		1.406	0.919	5	62.14	34.64
30		3.745	1.086	9	55.25	71.00

Table 5.7: Average waiting time under a deterministic bad weather scenario, under the best fixed allocation of $(N_A, N_B) = (7, 3)$, the original schedule allocation from the fair weather deterministic model and a new schedule allocation computed for this particular scenario, when decision epochs occur every thirty minutes; the total number of servers switched under the new schedule allocation; and the percentage improvement of the new schedule allocation over the fixed and original schedule allocations, by switching time θ.

allocation using the revised entry rates discussed in Section 5.3.2. First, we see that in general, the average waiting times are lower in this case where entry rates are shifted due to bad weather than in the fair-weather model. This unexpected result is easily explained: in this scenario, passenger entries are dispersed over a period of a few hours, distributing more evenly the load over the day. This is especially pronounced during typical peak periods, because the number of passengers shifted away from the peak period is greater than the number shifted into the peak period from non-peak periods. Thus, the peaks, which had caused most of the delays in the previous model, are now flatter, causing a disproportionate reduction in the average waiting time.

We note next that the fixed allocation for this case is once again (7,3), which is reasonable because the total expected number of customers entering Terminals C and E has not changed, only the distribution of these entrances over time. What is interesting, however, is that this fixed allocation can actually outperform the schedule allocation obtained from the original entry rates for certain values of θ. This suggests that in the face of widespread disruptions, relying on the same server allocation schedule as on a normal day can be worse than not allowing the server allocation to vary at all over the day. For instance, when $\theta = 0$, the original allocation mandates 18 switches over the course of the day, fluctuating to accommodate minor changes in entry rates. If the entry rates change, however, these fluctuations no longer match the new entry rates. And because queueing delays tend to propagate, even a one-period misallocation can cause large waiting times to be incurred over several hours. (As we see later in the chapter for the other model variants, the original schedule allocation for $\theta = 0$ is not a very robust allocation in that by being sensitive to queue imbalances in the original system, it is unable to do well if the entry pattern changes). When $\theta = 30$, we also see that the original schedule allocation fares worse than a fixed allocation. Here, only nine switches took place in the original allocation, so the large average waiting time is due not so much to fluctuations in the allocation but instead to large queues that form at the

θ	Fixed Allocation	Original Allocation	DP Allocation ($\tau = 30$)	% Imprvmt.	Expected Number of Switches
0	4.521 (7,3)	2.865	2.284	20.3	18.79
5		2.611	2.582	1.1	7.33
10		2.739	2.741	-0.1	7.54
15		3.013	2.902	3.7	8.18
30		3.707	3.420	7.7	8.10

Table 5.8: Average waiting time (minutes per customer) under a stochastically occurring bad weather scenario ($p = 1/15$), using the best fixed allocation of $(N_A, N_B) = (7, 3)$, the original schedule allocation, and a dynamic allocation; the percentage improvement of the dynamic allocation's waiting time over that of the original allocation; and the expected number of servers switched, by switching time, θ

start of the day and propagate. In our bad weather model, 30% of customers enter earlier than their scheduled period. In the first hour of the day, the original schedule allocation for $\theta = 30$ does not have enough servers on-duty at Terminal C to accommodate the customers who decide to enter early. Because it takes thirty minutes for any switched servers to come on-duty, queues that begin to form in the first period grow even longer while the system waits for servers to be reallocated.

The schedule allocation for the revised entry pattern yields significantly lower waiting times than either the original schedule allocation or the fixed allocation. Because misallocations can cause such large queueing delays, if a systematic change in the passenger entry pattern is anticipated, it is better to recompute an appropriate schedule allocation for that new pattern rather than relying upon the original schedule.

5.6.3 Stochastically disrupted entries

A more interesting case to consider is that described in Section 5.3.3 in which a bad weather watch is announced in a period with probability p and, once announced, affects the customer entry rate to the airport terminals over the next four hours. Such unexpected shifts in the passenger entry rates are more realistic, and the stochastic nature of this problem allows us to explore the potential advantages of dynamic allocations over schedule allocations. Table 5.8 shows the average waiting times incurred under a fixed allocation, the original schedule allocation from the deterministic case and the dynamic allocation of Equation (5.9), as well as the percentage improvement of the dynamic allocation over the original schedule and the expected number of switches that take place in the dynamic allocation, as a function of the switching time θ, for $p = 1/15$.

First we see that the average waiting times in this stochastic bad weather situation are significantly higher than those in the deterministic bad weather situation and also higher than the average waiting times under the original deterministic fair weather case. Recall that in the deterministic bad weather scenario, the reduction in waiting time was achieved

by the spreading out of passenger entries throughout the day. In this stochastic scenario, we assume that the bad weather, if it occurs, can affect up to four hours only, limiting the smoothing effects of dispersing passenger entries. Furthermore, the uncertainty around when the bad weather will start (and thus when servers might need to be reassigned) also naturally causes the optimal waiting time to increase.

We note that the original schedule allocation obtained for $\theta = 0$ did not perform well in this stochastic scenario, as it yielded a higher average waiting time than some allocations having non-zero switching times. Once again, this indicates a lack of robustness in the allocation, where it fails to perform well when applied to a new problem instance with moderately different entry rates. By contrast, the allocations obtained under non-zero switching times are characterized by a small number of switches occurring in response to important shifts rather than minor fluctuations in the customer entry pattern. Because these major shifts might only be somewhat affected by bad weather disruptions (for instance, a peak might be smoothed but will still remain a peak), allocations that respond only to major shifts in entry patterns but ignore slight fluctuations might be more robust to stochastic disruptions.

As a result of the original schedule's poor performance when $\theta = 0$, we see that a dynamically determined allocation offers a 20% reduction in the average waiting time. In cases with moderately larger switching times, the reduction is minimal because the original schedule allocation, to which the dynamic allocation is compared, is not as bad. As the switching times become even larger, the reduction in waiting time offered by a dynamic allocation increases. This suggests that when switching times are large, it is very important that switches be timed properly. The expected numbers of switches in the dynamic allocations are comparable to those of the schedule allocations, suggesting that dynamic allocations change only the timing of switches to achieve reductions in the average wait. However, these reductions are not guaranteed. When $\theta = 15$ minutes, the approximations made to solve the dynamic program resulted in the dynamic allocation actually performing somewhat worse than the original schedule allocation.

Figure 5-10 shows the histogram of deviations (in numbers of servers at Terminal C) between the dynamic allocation and the original schedule allocation for the periods prior to a bad weather watch, during a bad weather watch and after a bad weather watch, for small, medium and large switching times ($\theta = 0, 15, 30$). The frequency indicated is the number of states visited by the model having a particular deviation, and is not weighted by the likelihood that a state is visited.

We see that the dynamic allocation typically corresponds very closely to the original schedule allocation prior to the announcement of bad weather. Though we might expect the allocations to be identical during this period because the entry rates are unchanged prior to the start of bad weather, the dynamic program anticipates possible future disruptions to the entry pattern, causing its allocation to differ sometimes from the original schedule allocation. Nonetheless, in the majority of system states visited prior to the start of bad weather, anticipating the possibility of future bad weather yielded no difference from the original allocation.

The dynamic allocation differs most from the original schedule during the bad weather

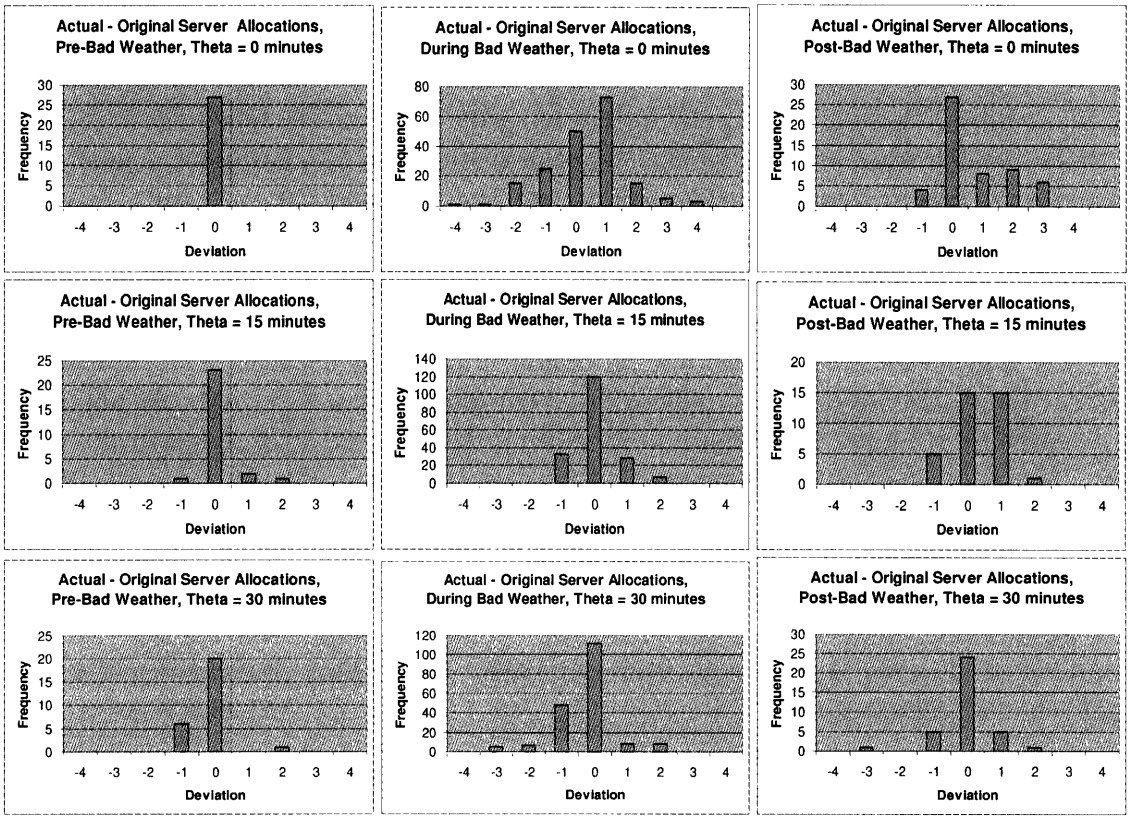

Figure 5-10: Difference between actual dynamic allocation and original schedule allocation to Terminal C in the stochastically disrupted entry model, before, during and after a bad weather watch for switching times of $\tau = 0$, 15 and 30 minutes.

θ	Fixed Allocation	Original Allocation	DP Allocation ($\tau = 30$)	% Imprvmt.	Expected Number of Switches
0	4.760 (7,3)	3.32	2.56	22.3	18.76
5		2.91	2.87	1.4	7.24
10		3.12	3.03	2.2	6.90
15		3.32	3.19	3.9	7.92
30		4.08	3.68	9.8	7.22

Table 5.9: Average waiting time (minutes per customer) under a stochastically occurring bad weather scenario ($p = 1/7$), using the best fixed allocation of $(N_A, N_B) = (7, 3)$, the original schedule allocation, and a dynamic allocation; the percentage improvement of the dynamic allocation's waiting time over that of the original allocation; and the expected number of servers switched, by switching time, θ

and after it has finished, with the distribution of deviations centered roughly around zero. When $\theta = 0$, we see that the dynamic allocation deviates more from the original allocation than when $\theta = 15$ or 30, which is not surprising given the significant improvement achieved by the dynamic allocation in that instance.

The value p corresponds to the likelihood bad weather is announced in any given period. Taking $p = 1/15$, the average time until bad weather is announced is 15 periods. This corresponds to noontime in our data, where neither queue is particularly busy. If we choose a different value of p, this could cause the bad weather to be more likely to be announced during a period of peak traffic at the security checkpoints, and might yield different results than $p = 1/15$. We tested the dynamic allocation for the case of $p = 1/7$, the results of which are shown in Table 5.9 We see, first, that the average waiting time is higher here than if $p = 1/15$, which we would expect because the bad weather is more likely to occur during busy periods. However, the percentage reduction in waiting time offered by the dynamic allocation is only somewhat greater.

In general, we conclude that the overall benefit of using a dynamic allocation in this situation where customer entry rates are affected by stochastic disruptions is minimal if the original schedule allocation is relatively stable. If the original allocation was very responsive to imbalances in the original system (e.g. when $\theta = 0$), there is greater benefit to altering the allocation to accommodate new disruptions. The charts shown in Figure 5-10 emphasize that much of the need for reallocation of servers occurs during or in the wake of bad weather, and that taking into account future bad weather before it starts has little influence over the server allocation. It is possible that stochastic disruptions of a greater magnitude or affecting the two queues unevenly might show greater benefits of dynamic allocation.

5.6.4 Randomized entry rates

In Section 5.3.4, we formulated a dynamic program for a system in which at each period and for each checkpoint independently, the arrival rate was a factor $(1 - \alpha)$ or $(1 + \alpha)$

	$\alpha = 0.1, \beta = 0.1$				$\alpha = 0.25, \beta = 0.25$			
θ	Fixed Alloc.	Original Alloc.	Dynamic Alloc	% Imprvmt.	Fixed Alloc.	Original Alloc.	Dynamic Alloc.	% Imprvmt.
0		2.05	1.94	5.4		4.34	3.17	27.0
5	4.24	2.27	2.25	0.9	5.41	3.80	3.53	7.1
10	(7,3)	2.46	2.42	1.6	(7,3)	4.00	3.72	7.0
15		2.64	2.59	1.9		4.22	3.92	7.1
30		3.18	3.21	-0.9		4.74	4.51	4.8

Table 5.10: Average waiting time (minutes per customer) when entry rates are randomized to equal $(1 - \alpha)\lambda_{\cdot}(t)$ or $(1 + \alpha)\lambda_{\cdot}(t)$ with probability β, using the best fixed allocation of $(N_A, N_B) = (7, 3)$, the original schedule allocation, and a dynamic allocation; and the percentage improvement of the dynamic allocation's waiting time over that of the original allocation, by switching time, θ, for $(\alpha = 0.1, \beta = 0.1)$ and $(\alpha = 0.25, \beta = 0.25)$

lower or higher than the original arrival rate $\lambda_{\cdot}(t)$ with probability β. Table 5.10 shows the average waiting time in such a system under fixed, schedule and dynamic allocations, for the cases of $(\alpha = 0.1, \beta = 0.1)$ and $(\alpha = 0.25, \beta = 0.25)$. We see that in the system where the actual entry rates are very close to the expected entry rates $(\alpha = 0.1, \beta = 0.1)$, dynamic allocation offers little reduction in the average waiting time over the original schedule allocation. However, in a system with a high level of uncertainty about the actual entry rates $(\alpha = 0.25, \beta = 0.25)$, dynamic allocation yields at least a 5% reduction in the average waiting time over the original schedule allocation. This suggests that the poorer the quality of information about entry rates, the more the system can benefit from determining allocations dynamically rather than relying upon a pre-determined schedule.

5.6.5 Stochastic service times

To explore the benefits of dynamic server allocation in the presence of stochastic service times, we use the two approximate dynamic programming heuristics described in Section 5.4.2: a *hybrid allocation*, in which the system evolves stochastically but allocations are decided assuming deterministic evolution, and a *nearest neighbor allocation*, in which the system evolves stochastically and allocations are determined based on this stochasticity but are restricted to a set of "nearest neighbors" of the original schedule. We compare, in Table 5.11, the performance of these heuristics to a fixed allocation and to the original schedule allocation applied to this stochastic instance.

We see first that the best fixed allocation is again to assign seven servers to Terminal C and three servers to Terminal E, and in this case it yields an average waiting time of 4.438 minutes per customer, an increase of about twenty seconds over the deterministic case, which indicates a "price of stochasticity". If we compare the original schedule allocation's performance in this stochastic case to the deterministic case shown in Table 5.4, we find that this price of stochasticity is relatively stable across the different switching times at around

θ	Fixed Alloc.	Original Schedule	Hybrid DP	% Imprvmt.	Avg. Switches	Nearest Neighbor	% Imprvmt.	Avg. Switches
0	4.438 (7,3)	2.357	**2.297**	2.5%	20.1	**2.213**	6.1%	20.9
5		2.490	**2.552**	-2.5	10.6	2.496	-0.2	16.1
10		2.810	**2.723**	3.1	9.7	**2.692**	4.2	12.1
15		2.943	2.904	1.3	10.0	**2.891**	1.8	12.3
30		3.457	3.448	0.3	8.76	3.418	1.1	11.1

Table 5.11: Average waiting time (minutes per customer) under a fixed allocation, the original schedule from the deterministic model, the hybrid dynamic allocation and the nearest neighbor dynamic allocation, as well as the average number of switches performed under the two dynamic programs, by switching time θ. Boldface entries indicate a statistically significant difference from the original schedule at the $\alpha = 0.05$ level.

25-30 seconds per customer.

We see further that while the dynamic allocations occasionally yield a statistically significant decrease in the average waiting time, the *practical* significance of such reduction is limited. In the best simulation (occurring when $\theta = 10$), the hybrid DP yielded only a 3.1% improvement over the schedule allocation, a difference of around five seconds per customer. In the worst case ($\theta = 5$), it performed 2.5%, or roughly four seconds, worse. The nearest neighbor DP performed only slightly better, but still did not offer consistent improvement over a schedule allocation. This suggests that one might be able to get an adequate solution using only the deterministic allocation without altering the allocation dynamically to accommodate variable service times.

The original schedule allocation corresponding to $\theta = 0$ did not perform as poorly in this instance of stochastic entries as it did under disrupted passenger entry rates. In those cases, the original $\theta = 0$ allocation was not robust to changes in the entry rates. Here, where only the service rates are stochastic, the original allocation seems largely unaffected by the stochasticity. This suggests, first, that optimal server allocations are generally more sensitive to changes in customer entry rates than to service rate fluctuations. Second, stochasticity on the level of individual customers appears to have less influence on the optimal allocation than do stochastic disruptions on an aggregate level. The entry rate disruptions we considered in the previous section occurred over a large period of time, affecting all customers scheduled to enter during that period. Such widespread change to the entry rate should naturally cause significant increases in waiting times if an allocation is not robust to these changes. By contrast, when a system's uncertainty is at the level of individual customers' service times, stochastic fluctuations tend to more or less even out over a decision epoch, and the ability to reallocate servers to accommodate this variability is less beneficial. It is possible that systematic changes to service times spanning one or more decision epochs might have a greater influence on the optimal server allocation. However, it is difficult to construct a realistic situation in which service rates of all customers would change over an extended period of time during the day.

5.6.6 General observations

Because we would like queue managers to be able to decide quickly whether to order a switch of servers between queues, it would be useful to distill the optimal switching policy into simple rules of thumb based on system state characteristics.

We attempt this first through the use of logistic regression. Logistic regression derives an expression for the likelihood that the DP heuristic switches a given number of servers, as a function of the explanatory variables. It is appropriate for models in which the response variable is ordinal (here, the response variable is the number of servers switched to Terminal C from Terminal E). The logistic regression model's best allocation is that having the highest likelihood.

The explanatory variables to include in the model are the queue lengths, Q_C and Q_E, the log-ratio of queue lengths (with a correction factor to accommodate null queue lengths), the number of servers N_C initially allocated to Terminal C at the start of the period, the entry rates $\lambda_C(t)$ and $\lambda_E(t)$ and their ratio $\lambda_C(t)/\lambda_E(t)$, the entry rates in the next period $\lambda_C(t+\tau)$ and $\lambda_E(t+\tau)$ and their ratio, and the ratio $\frac{\lambda_C(t)/\lambda_E(t)}{\lambda_C(t+\tau)/\lambda_E(t+\tau)}$, which measures how the load balance between the two queues is expected to change in the next period. If n_C is the best new allocation to use at Terminal C, then our decision variable in the logistic model is the value $n_C - N_C$, that is, the number of servers that should be switched to Terminal C from Terminal E, where a negative value means that servers should be switched out of Terminal C.

The logistic model assigns coefficients, α_i to each of the continuous variables, x_i, above (e.g., those pertaining to fluid queue lengths and entry rates) and assigns a coefficient β_j for each possible value, j, of ordinal variables (e.g., N_C). From this, it finds a linear expression

$$l = \sum_i \alpha_i x_i + \sum_j I_{N_C=j} \left[\sum_{i \le j} \beta_i \right], \tag{5.15}$$

where $I_{N_C=j}$ equals one if the number of servers currently allocated to Terminal C is j, and zero otherwise. Using intercept coefficients $\gamma(k)$ that are non-decreasing in k, the cumulative probability that the new allocation at Terminal C is no more than $k-1$ higher than the current allocation is given by

$$\text{CumProb}(k-1) = \frac{1}{1+e^{-(\gamma(k)+l)}}, \tag{5.16}$$

and the total probability of selecting $n_C - N_C = k$ is given by

$$\text{Prob}(k) = \text{CumProb}(k) - \text{CumProb}(k-1). \tag{5.17}$$

The best allocation at each state is to switch k servers from Terminal E to Terminal C, where k is the value having the highest Prob(k). Parameters that cause l to increase cause the cumulative probabilities to increase, which generally favors smaller values of k (servers should be moved away from Terminal C). Parameters that cause l to decrease generally favor

Parameter	$\theta = 0$	$\theta = 5$	$\theta = 10$	$\theta = 15$	$\theta = 30$
Q_C	-0.063	-0.047	-0.048	-0.042	-0.057
Q_E	0.050	0.060	0.062	0.060	0.043
$\log(\frac{Q_C}{Q_E})$	.	-0.128	-0.234	-0.147	-0.188
$\lambda_C(t)$	-2.072	-1.477	-1.217	-1.223	-2.616
$\lambda_E(t)$	0.965	3.235	3.553	3.661	12.221
$\lambda_C(t+\tau)$	0.554	.	-0.527	-0.374	-0.787
$\lambda_E(t+\tau)$	-0.531	-0.536	.	.	.
$\lambda_C(t)/\lambda_E(t)$	-4.527	.	2.539	3.087	6.757
$\lambda_C(t+\tau)/\lambda_E(t+\tau)$	1.173	1.125	.	.	.
$\frac{\lambda_C(t)/\lambda_E(t)}{\lambda_C(t+\tau)/\lambda_E(t+\tau)}$	5.931	3.517	.	.	11.734

Table 5.12: Coefficients from the logistic regression of change in server allocation ($n_C - N_C$) on system state characteristics, for variables that are significant at the $\alpha = 0.05$ level. Not shown: intercept coefficients and ordinal coefficients for the variables representing the initial server allocation. See Appendix B for complete regression analysis.

an increased number of servers at Terminal C. The significant (at $\alpha = 0.05$ level) continuous parameters and their estimated coefficients are shown in Table 5.12 for each of the switching times considered. (A (.) indicates that the parameter was not statistically significant in the model corresponding to that column's switching time). The complete regression output for the five models is given in Appendix B. We see that the coefficients for Q_C and Q_E have opposite signs, which we would expect, and that an increase in Q_C, which causes l to decrease, tends to favor an increased server allocation at Terminal C, and an increase in Q_E favors an increased allocation at Terminal E, which we would also expect. The magnitude of these two effects is roughly equal and unaffected by the switching time τ. The effects of the arrival rates $\lambda_C(t)$ and $\lambda_E(t)$ are also consistent in sign with what we might expect, but vary in magnitude for different values of θ. Other parameters tend to vary in significance level, magnitude and even sign across the different switching times. This makes it difficult to discern general rules about when switches should take place based on system state characteristics.

Nonetheless, even if general conclusions cannot be drawn from the logistic regression, it is interesting to explore whether the allocations recommended by the logistic regression model constitute a good server allocation policy. That is, rather than solving the dynamic program to determine when switches should occur, could an airport rely solely on this logistic regression model? To answer this, we simulate the system using the logistic regression model solution at each state for the server allocation. The average waiting times are summarized in Table 5.13, along with the original results from Table 5.11 for the stochastic service times model. We see that the logistic regression model performs roughly as well as the two dynamic allocations and the schedule allocation when tested on the same "training" data from which the model was created. However, we would like to know whether the relationship

θ	Fixed Allocation	Original Schedule	Hybrid DP	Nearest Neighbor	Logistic Regression
0	4.438 (7,3)	2.357	2.297	2.213	2.448
5		2.490	2.552	2.496	2.542
10		2.810	2.723	2.692	2.680
15		2.943	2.904	2.891	2.911
30		3.457	3.448	3.418	3.397

Table 5.13: Average waiting time (minutes per customer) under the logistic regression model tested on the training set, by switching time θ, as compared to the fixed, original schedule, hybrid and nearest neighbor allocations of Table 5.11.

θ	Schedule Allocation (Stochastic)	Hybrid DP	Nearest Neighbor	Logistic Regression
0	0.997	0.972	0.986	1.174
5	1.214	1.202	1.151	1.340
10	1.313	1.276	1.318	1.425
15	1.361	1.402	1.351	2.197
30	1.629	1.578	1.561	3.346

Table 5.14: Average waiting time (minutes per customer) under the logistic regression model tested on a test data set having entry pattern equal to that of Section 5.3.2 and stochastic service times, versus Section 5.3.2's original schedule allocation under stochastic service times and the hybrid and nearest neighbor allocations, by switching time θ.

derived in the regression holds across other sets of data. If we were to consider a new queueing system, could the logistic regression model predict when servers should be switched between the queues?

The first pitfall is that the above logistic regression model is based on the initial assumption that there are $N = 10$ servers that can be allocated to the queues. If N is smaller than 10, then the logistic regression model might suggest an allocation that is infeasible. If N is greater than 10, then there would a range of feasible allocations that would never be considered by the regression model. So an immediate constraint is that a given logistic regression model can be applied only to a system having the same number of servers as the system that generated the model.

To test our model's performance, we consider Terminals C and E again but use the entry rates given in Section 5.3.2 for the case of deterministic disruptions to the entry process. Table 5.14 shows the average waiting times under the schedule allocation of Section 5.3.2 applied to stochastic service times, the hybrid and nearest neighbor dynamic allocations and the logistic regression model allocation.

Although the logistic regression allocation does comparably well for small switching times, as the switching time increases, it begins to perform significantly worse than the schedule and dynamic allocations. This indicates that there might be additional variables excluded from the regression model that help to explain switching behavior in cases where switching times are higher. Though the possibilities are endless, one likely guess would be that the allocation under high switching times depends not only on the next period's entry rates but on entry rates in later future periods. When switches require significant amounts of time to be completed, it might become more important to consider changes to the entry rate further into the future.

Another heuristic that we considered was a modification of Duenyas and Van Oyen's heuristic in [45] to accommodate multiple servers, variable arrival rates and decision epochs of duration τ. In their heuristic, they modify the traditional $c\mu$-rule in a single-server system so that the decision considers not the maximum possible service rate μ but the time average service rate that is realized over the next busy period (the time until the queue being switched to empties), given that any switches cause the server to be temporarily unavailable, reducing total service capacity. In our version, we select the allocation that maximizes the total service rate over the next *two* decision epochs[3], computed as follows.

If all servers remain in their current positions, then the time average service rate at queue i will be $n_i\mu$ if the servers will be unable to work off any existing queue or arriving fluid over the next two time periods. If the queue is completely worked off by the end of the two periods, then the time average service rate is $\frac{Q_i}{2\tau} + \frac{\lambda_i(t)+\lambda_i(t+\tau)}{2}$. If the queue is worked off during the first decision epoch but reappears during the second decision epoch, then the average service rate at queue i is $\frac{Q_i}{2\tau} + \frac{\lambda_i(t)+n\mu}{2}$. This is summarized below:

$$\text{Rate}_i = \begin{cases} n_i\mu & Q_i + \lambda_i(t)\tau > n_i\mu\tau, \text{ and } Q_i + \tau(\lambda_i(t) + \lambda_i(t+\tau)) > 2n_i\mu\tau \\ \frac{Q_i}{2\tau} + \frac{\lambda_i(t)+\lambda_i(t+\tau)}{2} & Q_i + \lambda_i(t)\tau > n_i\mu\tau, \text{ and } Q_i + \tau(\lambda_i(t) + \lambda_i(t+\tau)) \leq 2n_i\mu\tau \\ & \text{or, } Q_i + \lambda_i(t)\tau \leq n_i\mu\tau, \text{ and } \lambda_i(t+\tau) \leq n\mu \\ \frac{Q_i}{2\tau} + \frac{\lambda_i(t)+n\mu}{2} & Q_i + \lambda_i(t)\tau \leq n_i\mu\tau, \text{ and } \lambda_i(t+\tau) > n\mu. \end{cases} \tag{5.18}$$

On the other hand, if k servers are switched from queue A to queue B, then the average service rate at queue A is the same as above, with $n_i = n_A - k$. At queue B, the average service rate is $\mu(n_B + k) - \frac{k\mu\theta}{2\tau}$ if the queue is never worked off during the two decision epochs. If the initial queue is worked off and no additional queue forms during the two decision periods, then the average service rate is $\frac{Q_B}{2\tau} + \frac{\lambda_B(t)+\lambda_B(t+\tau)}{2}$. Lastly, the average service rate is equal to $\frac{Q_B}{2\tau} + \frac{\lambda_B(t)+(n_B+k)\mu}{2}$ if the queue is worked off during the first decision

[3]Had we considered only the next decision epoch, then under very large switching times, servers would almost never be switched. For instance, if the switching time, θ, is equal to the decision interval, τ, then switching a server would be equivalent to permanently losing a server, as he would not arrive at the new queue before the end of the current period. In addition to this, considering two decision epochs allows us to consider a change in entry rate at the next period.

θ	Schedule Alloc.	Nearest Neighbor	Number of Switches	Max. Rate Heuristic	Number of Switches	% Imprvmt.
0	2.357	2.213	20.9	2.500	10.0	-13.0
5	2.490	2.496	16.1	2.647	10.0	-6.0
10	2.810	2.692	12.1	2.804	10.1	-4.2
15	2.943	2.891	12.3	2.972	10.2	-2.8
30	3.457	3.418	11.1	3.609	10.4	-5.6

Table 5.15: Average waiting times (minutes per customer) under the maximum service rate heuristic, by switching time θ, as compared to the original schedule and dynamic (nearest neighbor) allocations of Table 5.11. Also given are the average number of switches under the dynamic and heuristic allocations, and the percentage improvement of the heuristic allocation over the dynamic allocation.

epoch but reappears during the second decision epoch. This is summarized below:

$$
\text{Rate}_B = \begin{cases}
\mu(n_B+k) - \frac{k\mu\theta}{2\tau} & \begin{cases} Q_B + \lambda_B(t)\tau > (n_B+k)\mu\tau - k\mu\theta, \text{ and} \\ Q_B + \tau(\lambda_B(t) + \lambda_B(t+\tau)) > 2(n_B+k)\mu\tau - k\mu\theta \end{cases} \\
\frac{Q_B}{2\tau} + \frac{\lambda_B(t)+\lambda_B(t+\tau)}{2} & \begin{cases} Q_B + \lambda_B(t)\theta \le n_B\mu\theta, \text{ and} \\ \lambda_B(t+\tau) \le (n_B+k)\mu \end{cases} \\
& \text{or, } \begin{cases} Q_B + \lambda_B(t)\theta > n_B\mu\theta, \text{ and} \\ \lambda_B(t+\tau) \le (n_B+k)\mu, \text{ and} \\ Q_B + \lambda_B(t)\tau \le (n_B+k)\mu\tau - k\mu\theta \end{cases} \\
& \text{or, } \begin{cases} Q_B + \lambda_B(t)\theta > n_B\mu\theta, \text{ and} \\ Q_B + \lambda_B(t)\tau > (n_B+k)\mu\tau - k\mu\theta, \text{ and} \\ Q_B + \tau(\lambda_B(t) + \lambda_B(t+\tau)) < 2(n_B+k)\mu\tau - k\mu\theta \end{cases} \\
\frac{Q_B}{2\tau} + \frac{\lambda_B(t)+(n_B+k)\mu}{2} & \begin{cases} Q_B + \lambda_B(t)\theta \le n_B\mu\theta, \text{ and} \\ \lambda_B(t+\tau) > (n_B+k)\mu \end{cases} \\
& \text{or, } \begin{cases} Q_B + \lambda_B(t)\theta > n_B\mu\theta, \text{ and} \\ \lambda_B(t+\tau) > (n_B+k)\mu, \text{ and} \\ Q_B + \lambda_B(t)\tau \le (n_B+k)\mu\tau - k\mu\theta \end{cases}
\end{cases}
\tag{5.19}
$$

The heuristic is to select the number of servers, k, to switch from A to B (where $k = 0$ refers to the case of no switching in Equation (5.18), and $k < 0$ refers to switching from B to A, where A is substituted for B in Equation(5.19)), that yields the maximum average service rate, $\text{Rate}_A + \text{Rate}_B$, over the next two periods. Table 5.15 shows the average waiting time and average number of switches under this heuristic as compared to those under the nearest neighbor dynamic programming algorithm and original schedule allocation, for the case of stochastic service times. We see not only that the heuristic performs worse than the dynamic allocation, but that it generally fares worse than even the original schedule allocation.

Because the relationship between each system state and the allocation selected at that state is complex, it is difficult to identify rules of thumb for when switches should occur that would simplify the dynamic allocation computations but still perform better than the original schedule allocation. For situations where dynamic server allocation is warranted, the dynamic program would be more reliable than the heuristics tested above, and for other situations, a pre-determined schedule allocation might be sufficient.

Despite our inability to characterize the decision as a simple function of queue lengths and arrival rates, however, we can still make a few observations. First, as explained earlier, the number of switches conducted does not always decrease as θ, the switching time, increases. When switches require a very long period of time, additional servers may need to be switched to accommodate queues that grow during that time.

It is also seen in the results that the DP's decision may occasionally be to switch servers *away* from a longer queue, to a shorter queue. This is in contradiction with many results found in the literature for systems with or without switching times, and is most likely due to the fact that the decisions made in our models are allowed to anticipate future changes in the entry rates at each queue. Although Queue A might be longer than Queue B in the current period, knowledge that Queue B's arrival rate might suddenly increase in the near future could support a decision to move servers *away* from Queue A.

Another phenomenon that contradicts results found in the literature for simpler systems is that even when one queue is empty and the other queue very long, the dynamic program might not decide to switch a server to the very long queue. In a stochastic context, this is known as *idling*, where servers remain idle at an empty queue while another queue is very long. In a fluid model, however, as long as the entry rate to a queue is nonzero, no idling is possible. When the entry rate is smaller than the total service capacity, the servers work continuously to process the fluid as it arrives. Thus, even though a queue may not form, the system is not empty and servers are not idle. This could explain why the dynamic program may elect, in some cases, to leave servers at an "empty" queue rather than switch them over to a busier one.

5.6.7 Sensitivity analysis

According to Susan F. Hallowell of the Transportation Security Administration, the airline industry has a saying, "If you've seen one airport, you've seen one airport" [4]. That is, airports are served by different sets of airlines, have different flight schedules and traffic patterns, and most importantly in the security context, different floorplans affecting how security checkpoints are arranged and operated. A dynamic server allocation would be infeasible at an airport in which all passengers pass through the same security checkpoint or in which security checkpoints are separated by great distances. Some airports have multiple checkpoints serving a same set of gates, so that airport customers can select which checkpoint they use based on which has the shorter lines. In such a system, customers themselves could balance queue loads, reducing the need to allocate dynamically the security screeners. Still, other airports might be better suited to dynamic security allocation than Logan Airport, for which we found dynamic allocation to offer at best only moderate improvement in waiting

Allocation	θ	$N = 9$	$N = 10$
Original Schedule	0	7.217	2.865
	5	7.100	2.611
	10	7.368	2.739
	15	7.573	3.013
	30	8.320	3.707
Dynamic	0	6.639	2.284
	5	7.070	2.582
	10	7.299	2.741
	15	7.522	2.902
	30	8.036	3.420

Table 5.16: Average waiting time (minutes per customer) in the stochastic entry pattern disruption model when the number of servers is reduced from $N = 10$ to $N = 9$, under the original schedule allocation and the dynamic allocation.

times. Recognizing this renders it difficult to make any specific statements about when dynamic allocations should be conducted, but we will discuss how the benefits of dynamic allocations are sensitive to characteristics of the system under study. We explore here whether dynamic allocations could be used to reduce staffing levels at checkpoints, and in the next section, we discuss the limits of our dynamic server allocation framework.

Using dynamic allocation to reduce the number of servers

In the previous results sections, we have focused on the extent to which different types of allocations (fixed, schedule or dynamic) affect average customer waiting times in parallel queues where the total number of servers, N, is kept constant. An equally valid question is whether staffing levels can be reduced via dynamic allocations while maintaining satisfactory levels of service for the queue customers. Indeed, in the postal service study of [24], it was found that dynamically allocating back room employees to the front room in response to shifts in queue length could yield a reduction of 1-2 servers in test cases in which 6-15 servers were originally required.

Specifically of interest is the question of whether, in a stochastic system, the waiting times using $N - 1$ servers under a *dynamic* allocation are comparable to those using N servers under a *schedule* allocation, when otherwise having only $N - 1$ servers would be insufficient. Does dynamic allocation permit us to use fewer servers than would a schedule allocation? Based on the results from the previous section, the initial hypothesis is that dynamic allocation does not reduce waiting times sufficiently as compared to a schedule allocation to permit a reduction in staffing. Tables 5.16 and 5.17 demonstrate this.

Table 5.16 shows the average waiting times in the stochastically disrupted entry pattern model when $N = 9$ and 10 and under both schedule and dynamic allocations. We see that the average waiting times when only $N = 9$ servers are on-duty are significantly worse than

Allocation	θ	$N = 9$	$N = 10$
Original Schedule	0	6.441	2.357
	5	6.576	2.490
	10	6.904	2.810
	15	7.065	2.943
	30	7.580	3.457
Dynamic (Hybrid)	0	6.391	2.297
	5	6.615	2.552
	10	6.873	2.723
	15	7.116	2.904
	30	7.542	3.448
Dynamic (Nearest (Neighbor)	0	6.374	2.213
	5	6.535	2.496
	10	6.836	2.692
	15	7.050	2.891
	30	7.520	3.418

Table 5.17: Average waiting time (minutes per customer) in the stochastic service rate model when the number of servers is reduced from $N = 10$ to $N = 9$, under the original schedule allocation and the dynamic allocations yielded by the hybrid and nearest neighbor heuristics.

when $N = 10$. Although the dynamic allocation offers moderate improvement over the original schedule allocation when $N = 9$, it is not sufficient to justify a reduction in staffing from $N = 10$. Table 5.17 shows similar results for the stochastic service rates model. Once again, neither of the dynamic allocations offers significant improvement over the schedule allocation even for a same value of $N = 9$, let alone as compared to a system having $N = 10$ servers available.

To explore whether another system having different characteristics might yield more optimistic results, we examine server allocations between security checkpoints B2 and B5 (both in Terminal B at Logan) for entry rates experienced on January 19, 2005 between 8:30 AM and 7:30 PM. Checkpoint B2 has a capacity of up to two checkpoint lanes and B5 up to six. On January 19, Logan had a time average of 4.7 lanes open between the two checkpoints, so we run the allocation heuristics using $N = 4$ and $N = 5$. Table 5.18 shows the average waiting times for the case of stochastic service rates. Although the waiting times achieved by $N = 4$ servers are satisfactory, and an airport might be willing to incur these waiting times to reduce the number of screening teams, we note that it is not due to the *dynamic* allocation that this is possible, as the dynamic allocation performs on par with the original deterministic schedule.

We conclude, therefore, that while dynamic allocations can occasionally provide moderate reductions in average waiting times, they are unlikely to provide opportunities to reduce

Allocation	θ	$N = 4$	$N = 5$
Original Schedule	0	1.152	0.263
	5	1.538	0.256
	10	1.778	0.256
	15	1.977	0.259
	30	2.596	0.269
Dynamic (Hybrid)	0	1.169	0.271
	5	1.506	0.255
	10	1.764	0.259
	15	1.962	0.258
	30	2.534	0.267
Dynamic (Nearest (Neighbor)	0	1.120	0.205
	5	1.481	0.212
	10	1.767	0.217
	15	1.959	0.219
	30	2.571	0.247

Table 5.18: Average waiting time (minutes per customer) in the stochastic service rate model applied to Checkpoints B2 and B5, for $N = 4$ and $N = 5$, under the original schedule allocation and the dynamic allocations yielded by the hybrid and nearest neighbor heuristics.

staffing levels significantly.

5.7 Model limitations

Because this work is among the first in a relatively unexplored topic of queueing, we do not claim that the modeling techniques used here are the only valid methods for studying this problem. Rather, we have made a first cut at a fairly difficult queueing problem, and in the process, we have identified new avenues of exploration that might reveal improved performance of dynamic allocations. For instance, this work has focused on minimizing the total waiting time a passenger spends in queue at the airport. One might similarly consider an allocation that maximizes the total utilization (or throughput) of servers over the day. However, an optimal solution to such objective functions might force some passengers to wait very long while most experience no wait. A different objective function might be to minimize the variance in waiting times or to minimize the probability that *any* passenger waits longer than, say, ten minutes. However, these nonlinear objective functions are fundamentally more difficult to model.

We also assumed that the decision to switch would occur at pre-specified decision epochs. While it is unrealistic to assume that airport checkpoint queues could be monitored continuously and switches prescribed at any time, a rigid assumption of decision epochs automatically limits the potential responsiveness dynamic allocations are intended to provide. One possible solution is to allow the employees themselves to decide whether or not to switch to a different queue once they become idle. This could allow switches to occur more frequently but introduces a few obstacles. First, the servers would need to know the length of queues elsewhere in the airport and expected future entry rates in order to make an optimal decision. Without such information, the server might accidentally switch away from a temporarily idle queue that is likely to see a sudden surge in the entry rate or choose not to switch when perhaps he should. Secondly, even with access to this information, we have seen that the decision of when to switch is not easily expressed as a simple function of the queue lengths and entry rates. The heuristics tested here proved unreliable in determining when switches should occur, so each server would likely need to solve a dynamic program to decide whether or not to switch. While such obstacles could be overcome through the use of pagers and computers, they also demonstrate that the framework used here, involving a centralized queue manager, is perhaps more realistic, if less flexible.

Next, the server allocation framework presented here, involving two parallel queues and a fixed pool of servers, is not appropriate for all systems. As an example, we consider two checkpoints within Terminal C: Checkpoint C2, having five possible lanes, and Checkpoint C3, having four possible lanes. The time-average number of servers allocated to these checkpoints on January 18, 2005 between 5:00 AM and 6:30 PM was 5.62 servers. Because our model requires the number of servers, N, to remain constant over the day, we might expect that choosing either $N = 5$ or $N = 6$ could result in reasonable average waiting times. However, even using $N = 6$ (a more costly, on average, staffing level than that used by Logan) yields significantly higher waiting times than we have seen previously (see Table 5.19).

θ	Schedule Allocation
0	5.958
5	6.517
10	7.090
15	7.385
30	8.064

Table 5.19: Average waiting times (minutes per customer) for the deterministic fluid model schedule allocation applied to Checkpoints C2 and C3, when $N = 6$.

These waiting times might not initially appear unacceptable, but these are average waiting times over the entire day, including long periods of time where passengers experience near zero delays. Much of the contribution to this average delay occurs during the morning peak period, when both queues grow to nearly 300 customers for certain values of θ. The reason for this lies in how the entry rate varies across the two queues over the day. Figure 5-11 shows the expected entry rates at checkpoints C2 and C3 over the day. We see that their peaks and valleys roughly coincide, so that when queue C2 needs additional servers, C3 also needs additional servers and vice versa. The net result is that there are periods of time, such as between 9 AM and 1:30 PM, when the system can function well with relatively few servers on-duty (and indeed, Logan allocated fewer servers at this time), while between 7 and 8:30 AM, for instance, the system needs 8-10 servers to accommodate the entries (and Logan allocated more servers at this time). Holding the value of N constant over the day, as our model does, forces the system to be overstaffed during off-peak periods and understaffed during peak periods. As there is no reward for being overstaffed and only penalties for being understaffed, we require a larger value of N to achieve reasonable average waiting times from our heuristics. Thus, our framework is too restrictive for queueing systems in which the two queues experience peaks and valleys simultaneously. In such systems, it would be better to vary the total number of servers, N, over the day rather than changing just the allocation of these N servers between the two queues.

Dynamic server allocations also do not work very well between queues having only a few security checkpoint lanes. When $N_{A_{max}}$ or $N_{B_{max}}$ is relatively small, then changing the allocation by even one server can dramatically affect the utilization ratio ($\frac{\lambda_i}{n_i\mu}$) at each queue. Because of this, dynamic reallocations are unlikely to occur often.

Lastly, most airports have more than two security checkpoints. An important extension to this work would be to consider switches between several queues. While this could provide additional flexibility to a dynamic allocation, it would also increase dramatically the complexity of the problem.

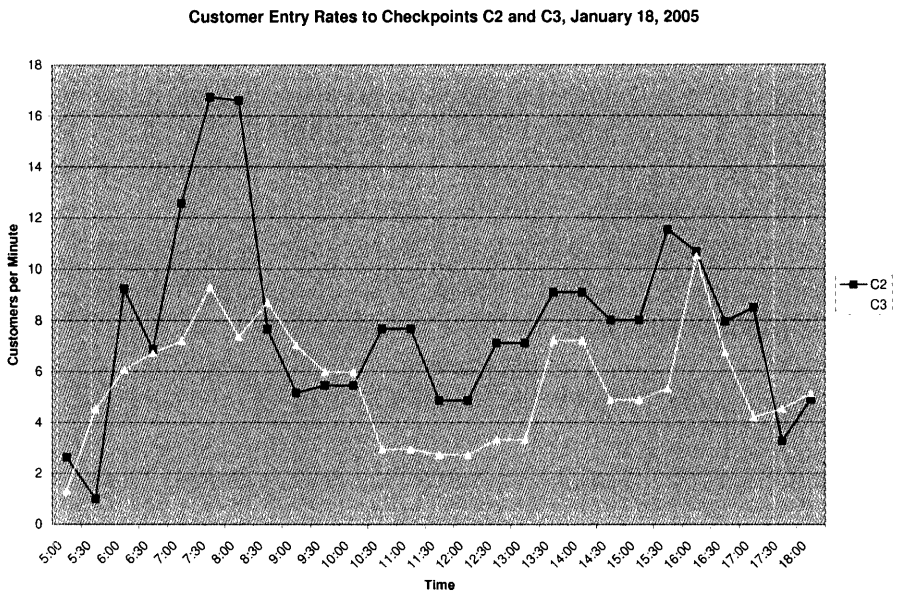

Figure 5-11: Entry rates (in customers per minute) to Checkpoints C2 and C3 at Logan Airport on January 18, 2005.

5.8 Are dynamic allocations of servers beneficial?

The models tested here have not provided evidence to support the dynamic allocation of servers at airports. Although it is clear that an anticipatory time-varying allocation, such as that provided by a deterministic schedule allocation, is superior to a fixed allocation, to take this time-sensitivity a step further and allow the allocation to depend on stochastic queue imbalances does not appear to significantly reduce waiting times in the examples considered.

One explanation is that the stochasticity we considered here was on a very small level. Stochastic service times, for instance, may cause some variability in queue lengths between decision epochs, but for the most part, the variability cancels itself out and the optimal server allocation is unaffected by this variability. Furthermore, in the presence of switching times, heuristics are reluctant to prescribe switches unless a major shift in customer entry pattern has occurred. Thus, the minor fluctuations in queue length caused by variable service or entry rates do not necessitate a dynamic response.

In the case of stochastic weather disruptions to the entry pattern, in which all passengers over a four-hour period were affected, the benefit of a dynamic allocation was more significant, and even more so in a system having randomized entry rates. This suggests that dynamically allocating servers might be more useful to counteract large stochastic effects (such as a gate-change in which a wave of passengers might need to pass through a different checkpoint than expected) on an occasional rather than continuous basis. Stochastic disruptions that affect the two queues unequally are also more likely to necessitate dynamic adjustments than disruptions that affect both queues proportionately. During normal queue operations,

the use of a schedule allocation based on deterministic entries and services appears to be sufficient.

In those cases where dynamic allocation could be useful, rules of thumb for when switches should occur are difficult to generalize and typically perform no better than the original schedule allocation. Fortunately, in these cases, the dynamic programming heuristics formulated in this chapter could be used to prescribe switches.

Certain system properties can also limit the benefit offered by dynamic allocations. For instance, queues in which the combined customer load varies widely over the day will not be well-served by a formulation where the total pool of servers is held fixed. And queues requiring only a few servers at any time generally wouldn't benefit from dynamic server allocation because temporarily losing even one server can cause the queue's load factor to skyrocket.

When airline passengers at the airport notice that one security queue is longer than another, a common reaction is to wonder why the airport does not reassign a screening team from the shorter queue to the longer one. Indeed some airports, such as Logan Airport in Boston and San Francisco Airport, do this. However, we have shown here that if the scheduled server allocation has been created based on accurate estimates of the customer entry rates to the respective queues, then dynamic reallocation of servers to compensate for stochastic fluctuations is not necessary unless those fluctuations are significant and affect customers on the aggregate, rather than individual, level.

Chapter 6

Conclusions

This thesis has examined some important problems in aviation security that have arisen in recent years. By relying on quantitative methods, we have tried to shed light on some of the intricacies of these issues, which are sometimes ignored in public debate. We began by quantifying the risk posed by aviation terrorism relative to that posed by other types of terrorism, and found that, even before 9/11, a randomly chosen American civilian was hundreds of times more likely to be killed by terrorists during an hour of an air journey than during an hour on the ground. While the Department of Homeland Security is often criticized for over-protecting aviation at the expense of other possible targets, we have demonstrated that terrorists have historically been fascinated by aviation and offered several examples to suggest that this fascination continues. Thus, an emphasis on aviation security is perhaps not simply fighting the last war but could be a reasonable reaction to a history of attacks.

We proceeded to examine specific security measures from the standpoints of detection capability, cost effectiveness and operational efficiency. A parameterized model for passenger pre-screening systems demonstrated that while public debate of such systems typically focuses on the criteria used for pre-screening, more important is the effectiveness of the actual screening performed on both general and high-risk passengers alike. When the profiling system is not robust to terrorist loopholes, such as their ability to probe the system prior to attacking, or when it is unable to draw ties between terrorists within a same group, it is far more important to ensure that the base level of screening experienced by all passengers be sufficient to thwart attacks. However, even when the profiling system is effective and robust, identifying high-risk passengers for additional screening is insufficient if the additional screening they receive is unable to detect a terrorist plot. Furthermore, our model also incorporated deterrence effects, by assuming that terrorists can gauge their likelihood of success prior to attacking and attack only if it is sufficiently high. While many critics cite terrorists' ability to probe the system as a major shortcoming of profiling systems, we have shown here that extensive probing could actually discourage terrorists from attacking, if the information they gain by probing reveals a lower chance of success than they initially anticipated. Although the lack of data and understanding of terrorist behavior impedes our ability to draw definite conclusions about profiling systems, our model can accommodate a wide range of assumptions and parameter values and it can serve as a tool for policy-makers

to evaluate the potential value of profiling systems.

We also investigated certain security measures from a cost standpoint, determining decision thresholds for the level of threat of attack required in order for a measure to be deemed cost-effective. The metric provided is a time threshold such that if an attempted attack is more likely than not to occur within this period of time, then the security measure should be implemented. Taking into consideration the costs of such measures, the costs of an attack and the cost and effectiveness of alternate security measures, we applied this model to three policies pertaining to luggage, airmail and cargo carried in the belly of passenger aircraft. We found that while the tendency might be to address all three issues as being identical, in fact, the policies have quite different costs and possibly different levels of threat, so that the cost-effectiveness of one does not necessarily imply the cost-effectiveness of the others. In particular, we find that even if attempts on checked luggage occur infrequently, matching checked bags to passengers on-board the aircraft is so inexpensive that it might still be cost-effective. On the other hand, unless an attempt to place an explosive inside of a cargo shipment is imminent, then removing cargo from passenger aircraft is not cost-effective, even if there exists no alternative screening method.

Finally, we considered an operational issue at passenger checkpoints at the airport, namely whether opening and closing lanes at security checkpoints across the airport dynamically, in response to fluctuations in queue lengths and arrival rates to the queue, can yield shorter waiting times than adhering to a pre-determined schedule for lane openings and closures. We used dynamic programming techniques on stochastic fluid models to determine optimal switching policies for systems having stochastic arrival patterns or stochastic service times. We found that, for the types of stochastic effects considered, dynamic allocation does not yield significant reductions in the average waiting time over a pre-determined schedule, despite its use at a few major U.S. airports. In particular, when the stochasticity affects passengers individually, such as in their service times, then the benefits of dynamic server allocation are negligible. However, when the stochasticity affects passengers on the aggregate, such as during weather disturbances in which the arrival rate experienced by all customers might shift or when the anticipated arrival rates are not very accurate, then dynamic allocation could be useful. In light of this, we explored, but were unable to attain, general rules guiding when such switches should occur. We found that switches were related in complex ways to the system state and that the use of general rules did not perform any better than a pre-determined schedule allocation and performed worse than the dynamic allocation heuristics we developed here.

The greatest limitation to these techniques is the lack of information. However, even absent key parameter values, we can still draw meaningful conclusions using these models. For instance, though we may not fully understand how terrorists are deterred by the presence of certain security measures, we can still observe the effects of deterrence assuming different levels of risk-averseness and show how the end assessment of the policy varies across these levels. Though we may not know the exact risk of an attack, we might believe that an attempt is more likely than not to occur over the next, say, twenty years. Given that looser statement of belief, a policy decision might defensibly be made.

These are just some of the issues facing aviation security policy, and there continue to be gaps that could be addressed using quantitative techniques. A common complaint against current efforts in aviation security is the lack of a system-wide perspective. While certain aspects of aviation security have been bolstered, others have been ignored, and policies appear to be implemented with little regard to how they interact with each other. For instance, much has been done to improve passenger security but little has been done to protect airport perimeters and aircraft access points. Said Dawn Deeks, spokeswoman for the Association of Flight Attendants, "It creates an illusion that we're doing everything we can. What passengers see is a very thorough search. We see behind the scenes at the airport, and the back door is wide open." [5]. Other concerns are threats to general aviation and commercial cargo aircraft.

Such a system-wide analysis should be a focus of future research, and the studies presented here are the building blocks of such work. Any large scale study would first need to evaluate the performance and costs of *individual* components of the security system prior to considering how they interact. The methodology presented here can help in this initial assessment and identify influential variables at the component level that might remain influential at the system level. Moreover, a systematic approach would require understanding how security measures interact to deter and thwart terrorism, which might involve extensive research over a long period of time. The frameworks given here applied to individual security measures could guide decisions in the meantime. Furthermore, the spirit of these techniques (incorporating terrorist behavior into a probabilistic assessment of a policy's performance, using the time until a first attack is likely to occur as a decision threshold for a policy's cost-effectiveness, and improving the operational performance of a security process) can also be applied to systems of security measures.

Future research must also keep up with the evolution of technology. Improvements to existing technology continue to be made and new technology for improving passenger and baggage screening continues to be developed, such as a "sniffer" capable of detecting traces of explosives on a person's body [4], radio frequency identification baggage tags and in-line baggage screening conveyor systems [32]. As faster screening technology becomes available, policies currently deemed too expensive or time-consuming might eventually be rendered cost-effective. Thus, aviation security decisions should not be considered to be one-time proclamations but should be continuously reviewed and updated to respond to new technologies and new threats. A key recommendation made by the 9/11 Commission was to "find a way of routinizing, even bureaucratizing, the exercise of imagination" [107] to conceive of new plots before they arise and to develop security policies accordingly.

As security policies are revised, a quantitative systems analysis approach should be used to model the interplay between security measures and guide security decisions. Even if aspects of terrorist deterrence or particular parameter values are not well understood, a broad sensitivity analysis can shed light on which measures might yield the greatest overall threat reduction, and which might simply cause the terrorists' attention to be diverted to other similar targets. While operational considerations, such as the efficient implementation of security procedures, are important and can also benefit from operations research techniques,

more important are the broader questions of which security policies to implement and what reduction in risk to expect from such policies. Some of the frameworks presented here could be used on future proposed security measures to evaluate their cost-effectiveness and their ability to thwart terrorism.

Appendix A

Fatal Terrorist Attacks Against American Civilians, 1/1/1968 - 9/10/2001

We list here fatal terrorist attacks against American civilians over the period 1/1/1968 to 9/10/2001, showing the number of civilians killed and the location category for the attack. A * denotes those attacks for which a location was not clearly stated in the source and that was assigned to a likely category based on similar types of attacks. *Sources: [1, 2, 17, 26, 101, 125, 145, 151, 152, 153].*

Table A.1: Fatal Terrorist Attacks

Date	Description	US Civilians Killed	Category
1-16-1968	Military advisor shot in car; Guatemala	2	Other Travel
8-28-1968	Ambassador taken/killed from car; Guatemala	1	Other Travel
10-12-1968	Army captain killed outside home; Brazil	1	Home
1-11-1970	Soldier killed in tavern; Ethiopia	1	Leisure
2-21-1970	Swissair Flight 330; Zurich	6	Aviation
2-23-1970	Tourist bus shooting; West Bank	1	Other Travel
6-10-1970	Murder of U.S. Army Assistant Attache at home	1	Home
7-31-1970	USAID worker kidnapped and killed	1	Home*
8-24-1970	Anti-Vietnam attack; U. of Wisconsin-Madison	1	Work
1-16-1972	Nurse killed in car; Gaza Strip	1	Other Travel
5-8-1972	Hijacking of Sabena Flight	1	Aviation
5-30-1972	Shooting at Lod Airport; Israel	16	Aviation
12-8-1972	Carbomb; Australia	1	Leisure
3-1-1973	Hostages at S. Arabia Embassy reception; Sudan	2	Leisure
5-21-1973	Killing of businessman; Argentina	1	Work
6-2-1973	American military advisor shot; Iran	1	Other Travel*
8-5-1973	Shooting of boarding flight passengers; Greece	2	Aviation
10-18-1973	Killing of hostage in bank robbery; Lebanon	1	Work
11-6-1973	Murder of school superintendent; California	1	Work
11-22-1973	Businessman and bodyguards shot; Argentina	3	Other Travel*
12-17-1973	Flight attacked at Fiumicino Airport; Italy	14	Aviation
8-6-1974	Bombing at Los Angeles Airport; California	2	Aviation
8-19-1974	US Ambassador killed at office; Cyprus	1	Work
9-8-1974	Bombing of TWA flight; Greece	17	Aviation
1-27-1975	Bombing of Fraunces Tavern; New York City	4	Leisure
2-26-1975	U.S. Consular agent kidnapped/killed; Argentina	1	Work*

Continued on next page

Table A.1: *Continued*

Date	Description	US Civilians Killed	Category
5-21-1975	USAF officers shot on way to work; Iran	2	Other Travel
12-23-1975	Shooting of diplomat outside home; Greece	1	Home
12-29-1975	Bombing at Laguardia Airport; New York City	11	Aviation
6-16-1976	Diplomats killed at roadblock; Lebanon	2	Other Travel
8-11-1976	Attack on passengers at Yesilkoy Airport; Turkey	1	Aviation
8-28-1976	Assassination of officials in car	3	Work
9-11-1976	Bomb at Grand Central Station; New York City	1	Other Travel
1-20-1977	Shooting of businessman; Mexico	1	Work
3-9-1977	Takeover of three buildings; Washington D.C.	1	Work
3-27-1977	Missionary killed by rebels; Ethiopia	1	Home
4-1-1977	Missionary killed by rebels; Zaire	1	Other Travel*
11-29-1977	Businessman killed; Indonesia	1	Other Travel*
12-2-1977	Businessman and bodyguards shot; Argentina	3	Other Travel
12-4-1977	Crash of hijacked aircraft; Malaysia	1	Aviation
6-3-1978	Bus bombing; Jerusalem	1	Other Travel
6-17-1978	Evangelist killed at work; Zimbabwe	1	Work
8-1-1978	Murder of policeman; Puerto Rico	1	Work
12-28-1978	Businessman shot in car; Iran	1	Other Travel
1-14-1979	Businessman killed in home; Iran	1	Home
2-14-1979	Ambassador kidnapped and killed; Afghanistan	1	Work*
4-12-1979	Two noncommissioned USAF officers shot; Turkey	4	Other Travel
4-30-1979	Killing of volunteer; Zimbabwe	1	Work
6-2-1979	Shooting of teacher at home; Turkey	1	Home
9-23-1979	Guerrilla attack on Instruction Center; El Salvador	3	Work
11-21-1979	Storming of US Embassy; Pakistan	2	Work
12-14-1979	Killing in front of government mini-bus stop; Turkey	4	Home
4-10-1980	Bombing of Airlines and Tourist Office; Turkey	1	Leisure
4-16-1980	Navy officer killed outside home; Turkey	1	Home
9-12-1980	Bombings; Philippines	1	Other Travel*
11-15-1980	Killing of two USAF officers at home; Turkey	2	Home
12-2-1980	Shooting of Catholic workers; El Salvador	4	Other Travel
12-7-1980	Businessman kidnapped and killed; Guatemala	1	Home
1-3-1981	Killing at restaurant; El Salvador	2	Leisure
1-19-1981	Linguist kidnapped and killed; Colombia	1	Work
2-1-1981	Kidnapping and killing of two tourists; Colombia	2	Leisure
5-16-1981	Bombing at JFK Airport; New York City	1	Aviation
6-23-1981	Kidnapping and killing of tourists; Zimbabwe	2	Leisure
7-27-1981	Priest shot and killed; Guatemala	1	Home
9-14-1981	Pastor shot and killed; Guatemala	1	Home
10-20-1981	Robbery at Nanuet Mall; New York	3	Work
1-18-1982	Assistant Military Attache shot outside home; France	1	Home
2-13-1982	Missionary shot and killed; Guatemala	1	Home
3-18-1982	Killing at plantation; Guatemala	1	Home
4-5-1982	Arson of restaurant; New York	1	Leisure
5-16-1982	Ambush of Navy sailors outside bar; Puerto Rico	1	Leisure
5-19-1982	Shooting; Puerto Rico	1	Leisure*
8-7-1982	Bombing and shooting at Ankara Airport; Turkey	1	Aviation
8-9-1982	Bombing of restaurant; France	2	Leisure
11-16-1982	Terrorist Robbery; Puerto Rico	1	Leisure
2-13-1983	Shooting; North Dakota	2	Work
4-18-1983	Bombing of US Embassy; Lebanon	17	Work
5-25-1983	US Military advisor killed in car; El Salvador	1	Other Travel
7-15-1983	Bombing of Orly Airport; France	2	Aviation
7-15-1983	Terrorist robbery of armored truck; Puerto Rico	1	Work
8-8-1983	Terrorist shooting; Michigan	1	Home
9-23-1983	Explosion of airlines; UAE	1	Aviation
11-15-1983	US Naval Mission chief killed in car; Greece	1	Other Travel
12-17-1983	Bombing at Harrod's Dept. Store; U.K.	1	Leisure
1-18-1984	Pres. of American U. of Beirut killed; Lebanon	1	Work
1-26-1984	Woman shot in car; El Salvador	1	Other Travel
2-15-1984	Force and Observer Dir. Gen'l murdered; Italy	1	Home

Continued on next page

Table A.1: *Continued*

Date	Description	US Civilians Killed	Category
4-12-1984	Restaurant bombing; Spain	18	Leisure
4-15-1984	Bombing at gas station; Namibia	2	Other Travel
6-18-1984	Shooting of Alan Berg; Colorado	1	Home
9-20-1984	Bombing of Embassy Annex; Lebanon	2	Work
11-30-1984	Librarian killed at Amer. U. of Beirut; Lebanon	1	Work*
12-4-1984	Hijacking of Kuwaiti flight; Iran	2	Aviation
3-13-1985	Bombing of restaurant; Guadeloupe	1	Leisure
5-12-1985	Terrorist murder of American; Philippines	1	Home
6-14-1985	Hijacking of TWA Flight 847; Lebanon	1	Aviation
6-19-1985	Shooting at café; El Salvador	13	Leisure
6-23-1985	Bombing of Air India flight; Ireland	19	Aviation
8-8-1985	Serviceman killed outside nightclub; West Germany	1	Leisure
8-15-1985	Bombing; New Jersey	1	Home
9-9-1985	Car bomb; Spain	1	Leisure
10-4-1985	Killing of kidnapped CIA officer; Lebanon	1	Other Travel*
10-7-1985	Hijacking of Achille Lauro cruiseship	1	Leisure
10-11-1985	Bombing; California	1	Work
11-23-1985	Hijacking of Eqyptair; Malta	1	Aviation
12-28-1985	Attack at Rome Airport; Italy	5	Aviation
4-2-1986	Bombing of TWA Flight 840; Greece	4	Aviation
4-5-1986	Bombing of nightclub; West Germany	2	Leisure
4-17-1986	Terrorist murders; Lebanon	1	Work*
4-29-1986	Terrorist shooting; Puerto Rico	1	Work*
5-17-1986	Death of kidnapped engineer; Colombia	1	Work*
6-25-1986	Bombing of tourist train; Peru	2	Other Travel
9-5-1986	Hijacking of Pan Am Flight 73; Pakistan	2	Aviation
10-25-1987	Killing of missionaries; Zimbabwe	2	Home*
6-13-1988	USAID subcontractor killed	1	Home*
6-28-1988	Car bomb of US Defense Attache; Greece	1	Other Travel
12-1-1988	Sabotage of tourist train; Peru	1	Other Travel
12-21-1988	Pan Am Flight 103; Scotland	189	Aviation
12-8-1988	Shooting of crop dusting plane; West Sahara	5	Work
3-15-1989	Political killing; El Salvador	1	Home*
5-24-1989	Shooting of missionaries; Bolivia	2	Home
7-6-1989	Terrorist bus takeover; Israel	1	Other Travel
8-18-1989	Shooting of teenage girl; West Bank	1	Home*
9-19-1989	Bombing of UTA Flight 772; Niger	7	Aviation
9-26-1989	Shooting of technicians in car; Philippines	2	Other Travel
11-11-1989	American killed by terrorist crossfire; El Salvador	1	Home*
11-21-1989	Terrorist murderer of reporter; Peru	1	Home*
11-30-1989	Shooting in pub; N. Ireland	1	Leisure
3-2-1990	Bombing of nightclub; Panama	1	Leisure
3-6-1990	Shooting of rancher; Philippines	1	Home
3-24-1990	Killing of missionary; Liberia	1	Home*
3-27-1990	Killing of missionary at home; Lebanon	1	Home
10-23-1990	Shooting of Iranian-born American; France	1	Home
2-7-1991	Shooting of contractor; Turkey	1	Home
3-12-1991	Bombing of apartment building; Greece	1	Home
3-22-1991	Shooting of contractor; Turkey	1	Work
10-28-1991	Bombing of parked vehicle; Turkey	1	Home
1-8-1992	Kidnapping and killing; Colombia	1	Leisure*
10-1-1992	Killing of kidnapped America; Colombia	1	Work*
10-20-1992	Attack on car; Liberia	2	Other Travel
1-31-1993	Kidnapping of missionaries; Colombia	3	Work
2-26-1993	Bombing of World Trade Center; New York City	6	Work
8-24-1993	Attack of student in car; South Africa	1	Other Travel
3-1-1994	Shooting of Jew on highway; New York	1	Other Travel
3-13-1994	Church shooting; South Africa	1	Leisure
7-19-1994	Bombing of Panamanian Atlas Flight; Panama	3	Aviation
10-9-1994	Kidnapping and killing; West Bank	1	Leisure*
1-15-1995	Attack on tourist's car; Cambodia	1	Leisure

Continued on next page

Table A.1: *Continued*

Date	Description	US Civilians Killed	Category
3-8-1995	Shooting of Consulate shuttle bus; Pakistan	2	Other Travel
4-9-1995	Suicide bombing; Gaza Strip	1	Other Travel
4-19-1995	Oklahoma City Bombing	168	Work
6-16-1995	Killing of kidnapped missionaries; Colombia	2	Leisure*
7-5-1995	Kidnapping-Murder India; trekker	1	Leisure
8-21-1995	Suicide bombing; Jerusalem	1	Other Travel
9-5-1995	Attack outside home; Jerusalem	1	Home
10-9-1995	Amtrak derailment; Arizona	1	Other Travel
11-13-1995	Car bombing at gov't office; Saudi Arabia	5	Work
11-19-1995	Shooting of UN worker; Bosnia-Herzegovina	1	Leisure*
2-25-1996	Suicide bombing on bus; Jerusalem	3	Other Travel
5-13-1996	Shooting near apartment complex; West Bank	1	Home
6-9-1996	Shooting on a car; Israel	1	Other Travel
7-27-1996	Centennial Olympic Park bombing; Georgia	1	Leisure
2-23-1997	Killing of kidnapped geologist; Colombia	1	Work
7-30-1997	Suicide bombing on outdoor market; Jerusalem	1	Leisure
9-4-1997	Suicide bombing at shopping mall; Jerusalem	1	Leisure
11-12-1997	Shooting of businessmen in vehicle; Pakistan	4	Other Travel
1-29-1998	Abortion Clinic bombing; Georgia	1	Work
8-7-1998	Kenya Embassy Bombing	12	Work
2-25-1999	Kidnapping and killing; Colombia	3	Work
3-1-1999	National Park killings; Uganda	2	Leisure
8-10-1999	Shootings; California	1	Work
5-25-2000	Journalist killed in ambush on vehicle; Sierra Leone	1	Other Travel
9-6-2000	Attack on UNHCR compound; Indonesia	1	Work
1-13-2001	Killing of kidnapped oil worker; Ecuador	1	Other Travel*
5-9-2001	Stoning of American teenager; West Bank	1	Leisure
5-27-2001	Kidnapping/Killing at resort; Philippines	1	Leisure
5-29-2001	American killed in car; West Bank	1	Other Travel
8-9-2001	Restaurant suicide bombing; Jerusalem	2	Leisure
7-2-1999-7-4-1999	Multiple Shootings: Illinois	2	Leisure

Appendix B

Logistic Regression of Dynamic Server Allocations

On the following pages we see the regression output for the final logistic regression models described in Section 5.6.6. The number of servers switched from Terminal E to Terminal C by the dynamic allocation is regressed on system variables such as queue lengths, entry rates, and functions thereof, for each value of θ. The coefficients for the variables $N_C[j-i]$ correspond to the coefficients β_j in the expression for l in equation (5.15) of Chapter 5 when the initial number of servers allocated to Terminal C, N_C, is equal to j.

For $\theta = 0$ minutes, we have the following expression for l (where the coefficients are obtained from Figure B-1):

$$\begin{aligned} l &= -0.06Q_C + 0.05Q_E - 2.07\lambda_C(t) + 0.96\lambda_E(t) \\ &\quad +0.55\lambda_E(t+\tau) + 5.93\frac{\lambda_C(t) * \lambda_E(t+\tau)}{\lambda_E(t) * \lambda_C(t+\tau)} - 4.53\frac{\lambda_C(t)}{\lambda_E(t)} + 1.17\frac{\lambda_C(t+\tau)}{\lambda_E(t+\tau)} \\ &\quad +\text{if } N_C = \begin{cases} 0,1,2 & : \quad 0 \\ 3 & : \quad 17.54 \\ 4 & : \quad 17.54 + (-2.73) = 14.80 \\ 5 & : \quad 14.80 + 14.86 = 29.66 \\ 6 & : \quad 29.66 + 10.01 = 39.67 \\ 7 & : \quad 39.67 + 10.13 = 49.80 \\ 8 & : \quad 49.80 + 8.00 = 57.80 \\ 9 & : \quad 57.80 + 8.92 = 66.72 \\ 10 & : \quad 66.72 + 7.49 = 74.21, \end{cases} \end{aligned}$$

and the following expressions for the cumulative probabilities:

$$\begin{aligned} Cumprob(-4) &= \frac{1}{1+e^{51.23-l}} \\ Cumprob(-3) &= \frac{1}{1+e^{42.86-l}} \\ Cumprob(-2) &= \frac{1}{1+e^{34.43-l}} \\ Cumprob(-1) &= \frac{1}{1+e^{23.66-l}} \\ Cumprob(0) &= \frac{1}{1+e^{14.43-l}} \\ Cumprob(1) &= \frac{1}{1+e^{2.34-l}} \\ Cumprob(2) &= \frac{1}{1+e^{-4.82-l}} \\ Cumprob(3) &= \frac{1}{1+e^{-6.67-l}} \\ Cumprob(4) &= 1 \end{aligned}$$

Ordinal Logistic Fit for Servers Switched, Theta = 0

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	1828.5006	17	3657.001	0.0000
Full	456.2884			
Reduced	2284.7890			

RSquare (U)	0.8003
Observations (or Sum Wgts)	1336

Converged by Objective

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	10663	456.28838	912.5768
Saturated	10680	0.00000	Prob>ChiSq
Fitted	17	456.28838	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[-3]	-51.229757	4.0972367	156.34	<.0001
Intercept[-2]	-42.859226	3.9008146	120.72	<.0001
Intercept[-1]	-34.436702	3.7362442	84.95	<.0001
Intercept[0]	-23.66549	3.524452	45.09	<.0001
Intercept[1]	-14.433052	3.296644	19.17	<.0001
Intercept[2]	-2.3433848	3.3247669	0.50	0.4809
Intercept[3]	4.82413226	3.0913492	2.44	0.1186
Intercept[4]	6.66982981	2.9898767	4.98	0.0257
Q_C	-0.0637495	0.0035012	331.53	<.0001
Q_E	0.05027258	0.0034559	211.61	<.0001
N_C[3-0]	17.5358507	2.5331746	47.92	<.0001
N_C[4-3]	-2.7315493	2.1687445	1.59	0.2078
N_C[5-4]	14.8596858	1.786541	69.18	<.0001
N_C[6-5]	10.0076901	0.6045313	274.05	<.0001
N_C[7-6]	10.1291601	0.6932371	213.49	<.0001
N_C[8-7]	8.0027656	0.5356627	223.20	<.0001
N_C[9-8]	8.92005749	0.745849	143.03	<.0001
N_C[10-9]	7.49404213	0.7723033	94.16	<.0001
λ_C(t)	-2.0720982	0.1397872	219.73	<.0001
λ_E(t)	0.96353281	0.2306196	17.46	<.0001
λ_C(t+τ)	0.55380564	0.0986257	31.53	<.0001
λ_E(t+τ)	-0.5309909	0.1684485	9.94	0.0016
λ_C(t)*λ_E(t+τ)/ λ_E(t)*λ_C(t+τ)	5.93132688	1.2805777	21.45	<.0001
λ_C(t)/λ_E(t)	-4.526884	0.588268	59.22	<.0001
λ_C(t+τ)/λ_E(t+τ)	1.17318598	0.4593916	6.52	0.0107

Figure B-1: Logistic Regression of Number of Servers Switched, $\theta = 0$

For $\theta = 5$ minutes, we have the following expression for l (where the coefficients are obtained from Figure B-2):

$$
\begin{aligned}
l \;=\; & -0.05Q_C + 0.06Q_E - 1.48\lambda_C(t) + 3.24\lambda_E(t) \\
& -0.54\lambda_E(t+\tau) - 0.13\log\left(\frac{Q_C}{Q_E}\right) + 1.12\frac{\lambda_C(t+\tau)}{\lambda_E(t+\tau)} + 3.52\frac{\lambda_C(t) * \lambda_E(t+\tau)}{\lambda_E(t) * \lambda_C(t+\tau)} \\
& +\text{if } N_C = \begin{cases} 0,1,2 & : \; 0 \\ 3 & : \; -7.54 \\ 4 & : \; -7.54 + 20.37 = 12.83 \\ 5 & : \; 12.83 + 16.51 = 29.34 \\ 6 & : \; 29.34 + 10.15 = 39.49 \\ 7 & : \; 39.49 + 11.13 = 50.62 \\ 8 & : \; 50.62 + 11.13 = 61.75, \end{cases}
\end{aligned}
$$

and the following expressions for the cumulative probabilities:

$$
\begin{aligned}
Cumprob(-2) &= \frac{1}{1+e^{65.13-l}} \\
Cumprob(-1) &= \frac{1}{1+e^{55.94-l}} \\
Cumprob(0) &= \frac{1}{1+e^{40.35-l}} \\
Cumprob(1) &= \frac{1}{1+e^{26.61-l}} \\
Cumprob(2) &= \frac{1}{1+e^{22.09-l}} \\
Cumprob(3) &= 1
\end{aligned}
$$

Ordinal Logistic Fit for Servers Switched, Theta = 5

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	1058.5270	14	2117.054	0.0000
Full	298.1917			
Reduced	1356.7187			

RSquare (U)	0.7802
Observations (or Sum Wgts)	1156

Converged by Objective

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	5761	298.19166	596.3833
Saturated	5775	0.00000	Prob>ChiSq
Fitted	14	298.19166	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[-1]	-65.134496	5186.7247	0.00	0.9900
Intercept[0]	-55.93979	5186.7245	0.00	0.9914
Intercept[1]	-40.347935	5186.7241	0.00	0.9938
Intercept[2]	-26.611485	5186.7241	0.00	0.9959
Intercept[3]	-22.086682	5186.7235	0.00	0.9966
Q_C	-0.0466762	0.003572	170.75	<.0001
Q_E	0.0595645	0.0051176	135.47	<.0001
N_C[3-0]	-7.5419657	6297.7869	0.00	0.9990
N_C[4-3]	20.3699701	3572.1182	0.00	0.9955
N_C[5-4]	16.5112984	1.735034	90.56	<.0001
N_C[6-5]	10.1478202	0.7111973	203.59	<.0001
N_C[7-6]	11.1294028	0.8819304	159.25	<.0001
N_C[8-7]	11.1308641	0.822756	183.03	<.0001
λ_C(t)	-1.4773548	0.1114318	175.77	<.0001
λ_E(t)	3.23547703	0.256	159.73	<.0001
λ_E(t+τ)	-0.5360498	0.1020765	27.58	<.0001
log(Q_C/Q_E)	-0.1275191	0.0358273	12.67	0.0004
λ_C(t+τ)/λ_E(t+τ)	1.12506193	0.5061198	4.94	0.0262
λ_C(t)/λ_E(t)*λ_E(t+τ)/λ_C(t+τ)	3.51724402	0.8838447	15.84	<.0001

Figure B-2: Logistic Regression of Number of Servers Switched, $\theta = 5$

For $\theta = 10$ minutes, we have the following expression for l (where the coefficients are obtained from Figure B-3):

$$\begin{aligned} l &= -0.05Q_C + 0.06Q_E - 1.22\lambda_C(t) + 3.55\lambda_E(t) \\ &\quad -0.53\lambda_C(t+\tau) + 2.54\frac{\lambda_C(t)}{\lambda_E(t)} - 0.23\log(\frac{Q_C}{Q_E}) \\ &\quad +\text{if } N_C = \begin{cases} 0,1,2,3 & : \quad 0 \\ 4 & : \quad 12.20 \\ 5 & : \quad 12.20 + 19.81 = 32.01 \\ 6 & : \quad 32.01 + 10.37 = 42.38 \\ 7 & : \quad 42.38 + 12.01 = 54.39 \\ 8 & : \quad 54.39 + 13.90 = 68.29, \end{cases} \end{aligned}$$

and the following expressions for the cumulative probabilities:

$$\begin{aligned} Cumprob(-2) &= \frac{1}{1+e^{73.36-l}} \\ Cumprob(-1) &= \frac{1}{1+e^{64.20-l}} \\ Cumprob(0) &= \frac{1}{1+e^{43.70-l}} \\ Cumprob(1) &= \frac{1}{1+e^{30.51-l}} \\ Cumprob(2) &= \frac{1}{1+e^{26.19-l}} \\ Cumprob(3) &= \frac{1}{1+e^{15.21-l}} \\ Cumprob(4) &= 1 \end{aligned}$$

Ordinal Logistic Fit for Servers Switched, Theta = 10

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	1130.0995	12	2260.199	0.0000
Full	314.1461			
Reduced	1444.2456			

RSquare (U)	0.7825
Observations (or Sum Wgts)	1283

Converged by Objective

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	7680	314.14615	628.2923
Saturated	7692	0.00000	Prob>ChiSq
Fitted	12	314.14615	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[-1]	-73.357841	229.04293	0.10	0.7488
Intercept[0]	-64.205187	229.03725	0.08	0.7792
Intercept[1]	-43.698745	229.02142	0.04	0.8487
Intercept[2]	-30.507077	229.02342	0.02	0.8940
Intercept[3]	-26.189666	229.04388	0.01	0.9090
Intercept[4]	-15.206751	225.49936	0.00	0.9462
Q_C	-0.0477533	0.0033012	209.25	<.0001
Q_E	0.06254654	0.0048716	164.84	<.0001
N_C[4-0]	12.1990331	229.00903	0.00	0.9575
N_C[5-4]	19.8087184	1.7821284	123.55	<.0001
N_C[6-5]	10.3698427	0.7076285	214.75	<.0001
N_C[7-6]	12.0129075	0.9023659	177.23	<.0001
N_C[8-7]	13.9028895	1.0471295	176.28	<.0001
λ_C(t)	-1.2170473	0.0943355	166.44	<.0001
λ_E(t)	3.55346168	0.2696832	173.62	<.0001
λ_C(t+τ)	-0.5274068	0.0732887	51.79	<.0001
logQ_C/Q_E	-0.2336195	0.0373829	39.05	<.0001
λ_C(t)/λ_E(t)	2.53784685	0.4572289	30.81	<.0001

Figure B-3: Logistic Regression of Number of Servers Switched, $\theta = 10$

For $\theta = 15$ minutes, we have the following expression for l (where the coefficients are obtained from Figure B-4):

$$\begin{aligned} l &= -0.04Q_C + 0.06Q_E - 1.22\lambda_C(t) + 3.66\lambda_E(t) \\ &\quad -0.37\lambda_C(t+\tau) + 3.09\frac{\lambda_C(t)}{\lambda_E(t)} - 0.15\log(\frac{Q_C}{Q_E}) \\ &\quad +\text{if } N_C = \begin{cases} 0,1,2,3 & : \quad 0 \\ 4 & : \quad 10.55 \\ 5 & : \quad 10.55 + 13.52 = 24.07 \\ 6 & : \quad 24.07 + 11.60 = 35.67 \\ 7 & : \quad 35.67 + 11.00 = 46.67 \\ 8 & : \quad 46.67 + 12.78 = 59.45, \end{cases} \end{aligned}$$

and the following expressions for the cumulative probabilities:

$$\begin{aligned} Cumprob(-2) &= \frac{1}{1+e^{70.22-l}} \\ Cumprob(-1) &= \frac{1}{1+e^{60.84-l}} \\ Cumprob(0) &= \frac{1}{1+e^{41.02-l}} \\ Cumprob(1) &= \frac{1}{1+e^{26.37-l}} \\ Cumprob(2) &= \frac{1}{1+e^{17.01-l}} \\ Cumprob(3) &= \frac{1}{1+e^{17.01-l}} \\ Cumprob(4) &= 1 \end{aligned}$$

Ordinal Logistic Fit for Servers Switched, Theta = 15

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	1030.6242	12	2061.248	0.0000
Full	299.2240			
Reduced	1329.8482			

RSquare (U)	0.7750
Observations (or Sum Wgts)	1292

Converged by Objective

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	6443	299.22400	598.448
Saturated	6455	0.00000	Prob>ChiSq
Fitted	12	299.22400	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[-1]	-70.22245	162.32381	0.19	0.6653
Intercept[0]	-60.845699	162.3165	0.14	0.7078
Intercept[1]	-41.024094	162.2986	0.06	0.8004
Intercept[2]	-26.369412	162.29553	0.03	0.8709
Intercept[4]	-17.009164	154.86194	0.01	0.9125
Q_C	-0.042322	0.0031334	182.43	<.0001
Q_E	0.06047581	0.0048496	155.51	<.0001
N_C[4-0]	10.5472906	162.28428	0.00	0.9482
N_C[5-4]	13.5257208	1.5086083	80.38	<.0001
N_C[6-5]	11.6046121	0.787972	216.89	<.0001
N_C[7-6]	11.0051471	0.857905	164.56	<.0001
N_C[8-7]	12.7750639	1.1801356	117.18	<.0001
λ_C(t)	-1.2236107	0.1016892	144.79	<.0001
λ_E(t)	3.66130822	0.2851798	164.83	<.0001
λ_C(t+τ)	-0.3742362	0.0682733	30.05	<.0001
log(Q_C/Q_E)	-0.1474304	0.0371215	15.77	<.0001
λ_C(t)/λ_E(t)	3.08700784	0.4845432	40.59	<.0001

Figure B-4: Logistic Regression of Number of Servers Switched, $\theta = 15$

For $\theta = 30$ minutes, we have the following expression for l (where the coefficients are obtained from Figure B-5):

$$\begin{aligned} l &= -0.06Q_C + 0.04Q_E - 2.62\lambda_C(t) + 12.22\lambda_E(t) \\ &\quad -0.79\lambda_C(t+\tau) + 6.76\frac{\lambda_C(t)}{\lambda_E(t)} - 0.19\log\left(\frac{Q_C}{Q_E}\right) + 11.73\frac{\lambda_C(t) * \lambda_E(t+\tau)}{\lambda_E(t) * \lambda_C(t+\tau)} \\ &\quad +\text{if } N_C = \begin{cases} 0,1,2,3 & : \quad 0 \\ 4 & : \quad 5.65 \\ 5 & : \quad 5.65 + 16.21 = 21.86 \\ 6 & : \quad 21.86 + 19.98 = 41.84 \\ 7 & : \quad 41.84 + 35.73 = 77.57 \\ 8 & : \quad 77.57 + 46.80 = 124.37, \end{cases} \end{aligned}$$

and the following expressions for the cumulative probabilities:

$$\begin{aligned} Cumprob(-2) &= \frac{1}{1+e^{181.08-l}} \\ Cumprob(-1) &= \frac{1}{1+e^{164.90-l}} \\ Cumprob(0) &= \frac{1}{1+e^{107.13-l}} \\ Cumprob(1) &= \frac{1}{1+e^{81.00-l}} \\ Cumprob(2) &= \frac{1}{1+e^{81.00-l}} \\ Cumprob(3) &= \frac{1}{1+e^{64.14-l}} \\ Cumprob(4) &= 1 \end{aligned}$$

Ordinal Logistic Fit for Servers Switched, Theta=30

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	926.5193	13	1853.039	0.0000
Full	195.6852			
Reduced	1122.2045			

RSquare (U)	0.8256
Observations (or Sum Wgts)	1250

Converged by Objective

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	6232	195.68520	391.3704
Saturated	6245	0.00000	Prob>ChiSq
Fitted	13	195.68520	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[-1]	-181.08013	290.85988	0.39	0.5336
Intercept[0]	-164.89572	290.77909	0.32	0.5707
Intercept[1]	-107.13248	290.58657	0.14	0.7124
Intercept[3]	-81.005845	287.34256	0.08	0.7780
Intercept[4]	-64.140488	269.49584	0.06	0.8119
Q_C	-0.0569997	0.0052079	119.79	<.0001
Q_E	0.04261333	0.0049356	74.54	<.0001
N_C[4-0]	5.6504764	281.86729	0.00	0.9840
N_C[5-4]	16.2119156	70.414781	0.05	0.8179
N_C[6-5]	19.9836521	2.4113066	68.68	<.0001
N_C[7-6]	35.7344952	4.427086	65.15	<.0001
N_C[8-7]	46.7950946	4.8638664	92.56	<.0001
λ_C(t)	-2.6160023	0.4151297	39.71	<.0001
λ_E(t)	12.221469	1.3357913	83.71	<.0001
λ_C(t+τ)	-0.7871487	0.1313489	35.91	<.0001
log(Q_C/Q_E)	-0.1877921	0.0518556	13.11	0.0003
λ_C(t)/λ_E(t)	6.75746241	1.3275267	25.91	<.0001
λ_C(t)λ_E(t+τ)/λ_E(t)λ_C(t+τ)	11.7343758	1.8204331	41.55	<.0001

Figure B-5: Logistic Regression of Number of Servers Switched, $\theta = 30$

Bibliography

[1] History of media and terrorism: Mediating terrorists. *Website.* http://terrorism.grady.uga.edu/history/1970s.html.

[2] Findings on the committee on government reform. *Puerto Rico Herald*, 3(53), 1999. Accessed at http://www.puertorico-herald.org/issues/vol3n53/ComiteFinalReport-en.shtml.

[3] Stowaway makes NY-Dallas flight in cargo crate. *Abilene Reporter News*, 9 Sept. 2003. Accessed at www.reporter-news.com/abil/nw_state/article/0,1874,ABIL_7974_2244898,00.html on 9 Sept. 2003.

[4] B. Adair. Building bombs to keep you safe. *St. Petersburg Times online*, 10 July 2002. Accessed at http://www.sptimes.com on 10 July 2003.

[5] S. Adcock. Airline ramp workers are never screened. *Newsday.com*, 8 Sept. 2003. Accessed at http://www.newsday.com on 9 Sept. 2003.

[6] Air Transport Association. Annual traffic and capacity: US airlines - scheduled. *Data Table*, 1926-2002. Accessed at http://www.airlines.org/econ/p.aspx?nid=1032 on 29 Oct. 2003.

[7] Air Transport Association. Airlines in Crisis: The Perfect Economic Storm. 2003. Accessed at http://www.airlines.org/econ/files/AirlinesInCrisis.pdf.

[8] R. Alonso-Zaldivar. Critics wary of new traveler profile system. *The Los Angeles Times online*, 26 Aug. 2003. Accessed at http://www.latimes.com on 2 Sept. 2003.

[9] S. Andradóttir, H. Ayhan, and D. G. Down. Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Science*, 47(10):1421–1439, Oct. 2001.

[10] S. Andradóttir, H. Ayhan, and D. G. Down. Dynamic server allocation for queueing networks with flexible servers. *Operations Research*, 51(6):952–968, Nov.-Dec. 2003.

[11] R. Anthony, B. Crane, and S. Hanson. Deterrence effects and Peru's Force-Down/Shoot-Down policy: Lessons learned for counter-cocaine interdiction operations. Paper P-3472, Institute for Defense Analysis, 2000.

[12] R. Anthony, B. Crane, and S. Hanson. The psychology of deterrence: A quantitative model. Research Summaries, Institute for Defense Analysis, 2000.

[13] G. E. Apostolakis. How useful is quantitative risk assessment? *Risk Analysis*, 24(3):515–520, 2004.

[14] M. Apprendi. Dogs of war protect Combined Joint Task Force - Horn of Africa. *Marine Corps News*, (200331291543), 11 Mar. 2003.

[15] Associated Press. American Airlines CEO urges some airport security measures be dropped. 31 May 2002.

[16] Associated Press. Cigarette lighters to be banned beyond security checkpoints. 3 Jan. 2005. Accessed at http://securityinfowatch.com/article/printer.jsp?id=2598 on 27 Feb. 2005.

[17] Aviation Safety Network. Safety issues website. Accessed at http://aviation-safety.net/events/index.html on 6 Mar. 2005.

[18] J. S. Baras, A. J. Dorsey, and A. M. Makowski. Two competing queues with linear costs: The μc rule is often optimal. *Advances in Applied Probability*, 17:186–209, 1985.

[19] A. Barnett. Aviation Security: A view from 2004. In *The Airline Industry*, Forthcoming.

[20] A. I. Barnett. CAPPS II: The foundation of aviation security? *Risk Analysis*, 24(4):909–916, 2004. Also Response to Discussants, pp. 933-934.

[21] A. I. Barnett, R. Shumsky, M. Hansen, A. R. Odoni, and G. Gosling. Safe at home? An experiment in domestic airline security. *Operations Research*, 49(2):181–195, Mar.-Apr. 2001.

[22] N. Bäuerle. Optimal control of queueing networks: An approach via fluid models. *Advances in Applied Probability*, 34:313–328, 2002.

[23] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *The Annals of Applied Probability*, 11(3):608–649, 2001.

[24] O. Berman and R. C. Larson. A queueing control model for retail services having back room operations and cross-trained workers. *Computers and Operations Research*, 31(2):201–222, Feb. 2004.

[25] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, Belmont, Massachusetts, second edition, 2000.

[26] P. Brogan. *World Conflicts*. Bloomsbury, London, 1998.

[27] C. Buyokkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Advances in Applied Probability*, 17:237–238, 1985.

[28] S. Chakrabarti and A. Strauss. Carnival Booth: An algorithm for defeating the Computer-Assisted Passenger Screening System. *First Monday*, 7(10), 2002. Available at http://www.firstmonday.org/issues/issue7_10/chakrabarti/index.html.

[29] D. W. Chen. New study puts September 11 payout at $38 Billion. *The New York Times*, page A1, 9 Nov. 2004.

[30] H. Chen and D. D. Yao. Dynamic scheduling of a multiclass fluid network. *Operations Research*, 41(6):1104–1115, Nov.-Dec. 1993.

[31] R.-R. Chen and S. Meyn. Value iteration and optimization of multiclass queueing networks. *Queueing Systems*, 32:65–97, 1999.

[32] A. Cho and T. Illia. Airports of the future: Next phase of baggage screening goes in-line, out of view. *McGraw-Hill Construction/ENR*. Accessed at http://enr.construction.com on 15 Jan. 2004.

[33] J. Chow, J. Chiesa, P. Dreyer, M. Eisman, T. W. Karasik, J. Kvitky, S. Lingel, D. Ochmanek, and C. Shirley. Protecting commercial aviation against the shoulder-fired missile threat. Occasional Paper 106, RAND Corporation, Santa Monica, CA, 2005.

[34] Close Quarter Battle K-9. Explosive detection K-9. Accessed at http://www.cqbk9.com in 2003.

[35] J. E. W. Cohen. Safety at what price? Setting anti-terrorist policies for checked luggage on US domestic aircraft. Master's Thesis, Massachusetts Institute of Technology, Sloan School of Management, Cambridge, MA, 2000.

[36] M. Coleman. Screening air cargo more complex than campaign ads suggest. *ABQ Journal*, 19 Oct. 2004. Accessed at http://www.abqjournal.com/elex/244456elex10-19-04.htm on 19 Oct. 2004.

[37] Committee on Science and Technology for Countering Terrorism, National Research Council. *Making the Nation Safer: The Role of Science and Technology in Countering Terrorism*. The National Academies Press, Washington, D.C., 2000. Available at http://www.nap.edu.

[38] C. C. Coughlin, J. P. Cohen, and S. R. Khan. Aviation security and terrorism: A review of the economic issues. Working Paper 2002-009 A, Federal Reserve Bank of St. Louis, 2002.

[39] C. A. Courcoubetis and M. I. Reiman. Optimal dynamic allocation of heterogeneous servers under the condition of total overload: the discounted case. In *Proceedings of the 27^{th} Conference on Decision and Control*, pages 634–639, Austin, TX, Dec. 1988.

[40] B. D. Crane, A. R. Rivolo, and G. C. Comfort. An empirical examination of counter-drug interdiction. Paper P-3219, Institute for Defense Analysis, Jan. 1997.

[41] P. K. Davis and B. M. Jenkins. Deterrence and influence in counterterrorism: A component in the war on al Qaeda. Monograph Report MR-1619, RAND Corporation, Santa Monica, CA, 2002.

[42] M. Derr. With training, a dog's nose almost always knows. *The New York Times*, 29 May 2001. Accessed at http://www.nytimes.com on 28 May 2003.

[43] M. Derr. With dog detectives, mistakes can happen. *The New York Times*, page F1, 24 Dec. 2002.

[44] L. Dixon. Compensation for losses from the 9/11 attacks. Monograph Report 264, RAND Corporation, Santa Monica, CA, 2004.

[45] I. Duenyas and M. P. Van Oyen. Heuristic scheduling of parallel heterogeneous queues with set-ups. *Management Science*, 42(6):814–829, June 1996.

[46] T. Eckert. Weapons screening at airports failing, report says. *The San Diego Union-Tribune, online*, 20 Apr. 2005. Accessed at http://signonsandiego.com on 25 Apr. 2005.

[47] D. Eggen. Airports screened nine of September 11 hijackers, Officials say kin of victims call for inquiry into revelation. *The Washington Post*, page A11, 2 Mar. 2002.

[48] B. Elias. Air cargo security. Report for Congress RL32022, Congressional Research Service, 11 Sept. 2003.

[49] E. Emery. Laser incidents worry pilots. *Denver Post*, 4 Jan. 2005.

[50] M. Fisher, C. Kubicek, P. McKee, I. Mitrani, J. Palmer, and R. Smith. Dynamic allocation of servers in a grid hosting environment. In *Proceedings of the fifth IEEE/ACM International Workshop on Grid Computing*, 2004.

[51] K. J. Garner. E-mail correspondence. *Auburn University*, 30 June 2003.

[52] K. J. Garner, L. Busbee, P. Cornwell, J. Edmonds, K. Mullins, K. Rader, J. M. Johnston, and J. M. Williams. Duty cycle of the detector dog: A baseline study. Institute for Biological Detection Systems, Auburn University, Apr. 2001.

[53] B. J. Garrick. Perspectives on the use of risk assessment to address terrorism. *Risk Analysis*, 22(3):421–423, 2002.

[54] A. Gathright. Airport anti-terror agency under fire for security gaps; critics cite billions spent, lapses in safeguards. *San Francisco Chronicle*, 9 Sept. 2003. Accessed at http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2003/09/09/MN261930.DTL on 10 Sept. 2003.

[55] L. V. Green and D. Guha. On the efficiency of imbalance in multi-facility multi-server service systems. *Management Science*, 41(1):179–187, Jan. 1995.

[56] R. E. Gunther. Server transfer delays in a dual resource constrained parallel queueing system. *Management Science*, 25(12):1245–1257, Dec. 1979.

[57] B. Hajek. Optimal control of two interacting service stations. *IEEE Transactions on Automated Control*, AC-29(6):491–499, June 1984.

[58] M. Hall. Airport screeners missed weapons. *USA Today*, 23 Sept. 2004. Accessed at http://www.usatoday.com/news/nation/20040922weapons_x.htm.

[59] C. Hallett. Remarks of Carol Hallett, President and CEO, Air Transport Association. *The International Aviation Club of Washington*, 18 June 2002.

[60] J. M. Harrison. The BIGSTEP approach to flow management in stochastic processing networks. In *Stochastic Networks: Theory and Applications*, pages 57–90. Oxford University Press, 1996. F. P. Kelly, S. Zachary and I. Ziedins, editors.

[61] J. M. Harrison and M. J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33:339–368, 1999.

[62] A. Harter. Security agency looks to profiling. Travelers privacy at risk, critics say. *Arkansas Democrat-Gazetteer*, page A1, 19 May 2002.

[63] J. Heimlich. E-mail correspondence. *Air Transport Association*, 2003.

[64] F. S. Hillier and K. C. So. On the simultaneous optimization of server and work allocations in production line systems with variable processing times. *Operations Research*, 44(3):435–443, May-Jun. 1996.

[65] M. Hofri and K. W. Ross. On the optimal control of two queues with server setup times and its analysis. *SIAM Journal of Computing*, 16:399–420, 1987.

[66] S. S. Hsu and B. Graham. Air security 'seriously flawed'. *The Washington Post*, page A22, 23 July 2004.

[67] C. Hulse. Threats and Responses: Airline Security, Lawmakers criticize Bush's air safety efforts. *The New York Times*, page A7, 9 Aug. 2003.

[68] S. Jacobson, J. L. Virta, J. M. Bowman, J. E. Kobza, and J. Nestor. Modeling aviation baggage screening security systems: A case study. *IIE Transactions*, 35(3):259–269, 2003.

[69] S. H. Jacobson, J. M. Bowman, and J. E. Kobza. Modeling and analyzing the performance of aviation security systems using baggage value performance measures. *IMA Journal of Management Mathematics*, 12:3–22, 2001.

[70] S. H. Jacobson, J. E. Kobza, and A. S. Easterling. A detection theoretic approach to modeling aviation security problems using the knapsack problem. *IIE Transactions*, 33:747–659, 2001.

[71] S. H. Jacobson, L. A. McLay, J. Kobza, and J. M. Bowman. Modeling and analyzing multiple station baggage screening security system performance. *Naval Research Logistics*, 52(1):30–45, 2005.

[72] O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, Oct. 1996.

[73] J. Jones. Airlines to test new security measures - American CEO. *Reuters*, 6 Sept. 2002.

[74] L. L. Jordan. Government report says aviation system still vulnerable. *SecurityInfoWatch.com*, 16 Mar. 2005. Accessed at http://www.securityinfowatch.com/article/printer.jsp?id=3361 on 25 Apr. 2005.

[75] G. A. Karayiannakis. Positive Bag-Match and September 11: Some dilemmas for aviation security. Master's Thesis, Massachusetts Institute of Technology, Sloan School of Management, Cambridge, MA, June 2002.

[76] S. Kehaulani Goo. Fliers to be rated for risk level. *The Washington Post online*, 9 Sept. 2003. Accessed at http://www.washingtonpost.com on 9 Sept. 2003.

[77] S. Kehaulani Goo. Hundreds of pilots trained to carry guns. *The Washington Post*, page A10, 27 Aug. 2003.

[78] S. Kehaulani Goo. US to push airlines for passenger records. *The Washington Post online*, 12 Jan. 2004. Accessed at http://www.washingtonpost.com on 14 Jan. 2004.

[79] P. T. Kilborn. Demand, public and private, booms for dogs to sniff out security threats. *The New York Times*, page A12, 12 Nov. 2001.

[80] J. Kobza and S. H. Jacobson. Addressing the dependency problem in access security system architecture design. *Risk Analysis*, 16(6):801–812, 1996.

[81] J. Kobza and S. H. Jacobson. Probability models for access security system architectures. *Journal of the Operational Research Society*, 48(3):255–263, 1997.

[82] D. Koenig. FBI probes man who shipped self to Dallas. *Associated Press*, 9 Sept. 2003.

[83] G. Koole. Assigning a single server to inhomogeneous queues with switching costs. *Theoretical Computer Science*, 182:203–216, 1997.

[84] G. Koole. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems*, 30:323–339, 1998.

[85] H. Kunreuther. The role of insurance in managing extreme events: Implications for terrorism coverage. *Risk Analysis*, 22(3):427–437, 2002.

[86] E. Lichtblau. Government report on U.S. aviation warns of security holes. *The New York Times online*, 14 Mar. 2005. Accessed at http://www.nytimes.com.

[87] E. Lipton. Chertoff restructuring homeland security department. *The New York Times online*, 13 July 2005. Accessed at http://www.nytimes.com on 18 July 2005.

[88] E. Lipton. Senate democrats assail domestic security budget. *The New York Times*, page A14, 13 July 2005. Accessed at http://www.nytimes.com on 18 July 2005.

[89] E. Lipton. Senators clash on security cost. *The New York Times online*, 15 July 2005. Accessed at http://www.nytimes.com on 18 July 2005.

[90] E. Lipton. Transportation security agency criticized. *The New York Times*, page A18, 19 Apr. 2005.

[91] Z. Liu, P. Nain, and D. Towsley. On optimal polling policies. *Queueing Systems*, 11:59–83, July 1992.

[92] R. Looney. Economic costs to the US stemming from the 9/11 attacks. Strategic Insight, Center for Contemporary Conflict, National Security Affairs Department, Naval Postgraduate School, 5 Aug. 2002.

[93] F. V. Lu and R. F. Serfozo. M/M/1 Queueing decision processes with monotone hysteretic policies. *Operations Research*, 32:1116–1132, 1984.

[94] A. Mandelbaum and A. L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research*, 52(6):836–855, Nov.-Dec. 2004.

[95] D. Marois. House passes homeland security conference report. *Aviation Now website*, 25 Sept. 2003. Accessed at http://www.aviationnow.com/avnow/news/channel_aviationdaily_story.jsp?id=news/hsb09253.xml.

[96] G. Martin. Identity crisis. *Condé Nast Traveler*, page 280, Nov. 2002.

[97] S. Martonosi and A. Barnett. Terror is in the air. *Chance*, 17(2):25–27, Spring 2004.

[98] A. J. Mason, D. M. Ryan, and D. M. Panton. Integrated simulation, heuristic and optimisation approaches to staff scheduling. *Operations Research*, 46(2):161–175, Mar.-Apr. 1998.

[99] P. McGeehan. Port Authority to improve airport security inspections. *The New York Times*, page B3, 25 Feb. 2005.

[100] L. A. McLay, S. H. Jacobson, and J. E. Kobza. Multilevel passenger screening strategies for aviation security systems. *Working Paper*, 2003.

[101] K. McMurray. *LaGuardiaBombing.com Website*, 1998. Accessed at http://www.laguardiabombing.com on 28 Oct. 2003.

[102] K. M. Mead. Challenges facing TSA in implementing the Aviation and Transportation Security Act. *Statement Before the Committee on Transportation and Infrastructure; Subcommittee on Aviation; United States House of Representatives*, 23 Jan. 2002. Report CC-2002-088.

[103] K. M. Mead. Key challenges facing the Transportation Security Administration. *Statement Before the Committee on Appropriations; Subcommittee on Transportation; United States House of Representatives*, 20 June 2002.

[104] S. P. Meyn. Stability and optimization of multiclass queueing networks and their fluid models. Lectures in Applied Mathematics, Volume 33, pages 175–199, Providence, RI, 1997. American Mathematical Society.

[105] J. J. Moder and C. R. Phillips, Jr. Queuing with fixed and variable channels. *Operations Research*, 10:218–31, 1962.

[106] L. F. Mullin. Prepared statement of Leo F. Mullin, Chairman and CEO, Delta Air Lines. *Before the Transportation and Infrastructure Committee, Aviation Subcommittee, U. S. House of Representatives Regarding the Financial Condition of the Airline Industry*, 24 Sept. 2002.

[107] National Commission on Terrorist Attacks Upon the United States. *The 9/11 Commission Report: Final Report.* Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., Official Government edition, 2004.

[108] G. F. Newell. *Applications of Queueing Theory.* Chapman and Hall, London, 1st edition, 1971.

[109] R. O'Harrow, Jr. Air security focusing on flier screening; Complex profiling network months behind schedule. *Washington Post*, page A1, 4 Sept. 2002.

[110] R. O'Harrow, Jr. Intricate screening of fliers in works. *The Washington Post online*, 31 Jan. 2002. Accessed at http://www.washingtonpost.com.

[111] C. Page. Term limits for security rules? Commentary. *The Washington Times online*, 3 Sept. 2003. Accessed at http://www.washtimes.com on 3 Sept. 2003.

[112] P. Page. Air cargo - Demand issues. *28th Annual FAA Aviation Forecast Conference*, 19 Mar. 2003. Transcript. Accessed at http://apo.faa.gov/Conference/2003/BreakPanel3/Page.htm on 1 July 2003.

[113] J. Palmer and I. Mitrani. Dynamic server allocation in heterogeneous clusters. Working paper, University of Newcastle upon Tyne, May 2003.

[114] J. Palmer and I. Mitrani. Optimal server allocation in reconfigurable clusters with multiple job types. In *Proceedings of the 2004 International Conference on Computational Science and its Applications, Part II*, LNCS 3044, pages 76–86, 2004.

[115] Parcel Shippers Association. USPS benefits from shifting parcels to FedEx planes. *PSA News*, 4 Oct. 2002.

[116] A. Ramirez. Golden noses are in demand, and they don't work just for food. *The New York Times online*, 12 Aug. 2004. Accessed at http://www.nytimes.com on 12 Aug. 2004.

[117] W. Rosa-Hatko and E. A. Gunn. Queues with switchover - a review and critique. *Annals of Operations Research*, 69:299–322, 1997.

[118] Salomon Smith Barney. FedEx Corporation (FDX) FDX:US Postal Contract As good as it gets. 29 Apr. 2002.

[119] G. Schneider. Terror risk cited for cargo carried on passenger jets; 2 reports list security gaps. *Washington Post*, page A1, 10 June 2002.

[120] P. Seidenstat. Terrorism, airport security and the private sector. *Review of Policy Research*, 21(3):275–291, May 2004.

[121] J. Sethuraman and M. S. Squillante. Optimal stochastic scheduling in multiclass parallel queues. In *Proceedings of ACM SIGMETRICS '99*, Atlanta, GA, May 1999.

[122] P. Shenon. US is inspecting overseas airports for missile threats. *The New York Times online*, 7 Aug. 2003. Accessed at http://www.nytimes.com on 7 Aug. 2003.

[123] R. Singel. CAPPS II stands alone, Feds say. *Wired News*, 13 Jan. 2004. Accessed at http://www.wired.com/news/privacy/0,1848,61891,00.html.

[124] V. K. Smith and D. G. Hallstrom. Designing benefit-cost analyses for homeland security policies. Working Paper, 11 Aug. 2004.

[125] J. Springer. LaGuardia Christmas bombing remains unsolved 27 years later. *CNN.com*, 24 Dec. 2002. Accessed at http://www.cnn.com/2002/LAW/12/24/ctv.laguardia/ on 28 Oct. 2003.

[126] M. S. Squillante, C. H. Xia, D. D. Yao, and L. Zhang. Threshold-based priority policies for parallel-server systems with affinity scheduling. In *Proceedings of the American Control Conference*, Arlington, VA, 25-27 June 2001.

[127] S. Stidham, Jr. and R. Weber. A survey of Markov decision models for control of networks of queues. *Queueing Systems*, 13:291–314, 1993.

[128] C. Strohm. Department of Homeland Security budget puts Administration on collision course with lawmakers, airlines. *GovExec.com Daily Briefing*, 9 Feb. 2005. Accessed at http://www.govexec.com/dailyfed/0205/020905c1.htm on 27 Feb. 2005.

[129] C. Strohm. TSA chief says aviation attacks still pose greatest risk. *GovExec.com Daily Briefing*, 15 Feb. 2005. Accessed at http://www.govexec.com/story_page.cfm?articleid=30563&printerfriendlyVers=1& on 7 Mar. 2005.

[130] D. Y. Sze. A queueing model for telephone operator staffing. *Operations Research*, 32(2):229–249, Mar.-Apr. 1984.

[131] L. Tanner. Airlines suffering from loss of mail revenue. *Dallas Business Journal*, 2 Dec. 2002. Accessed at http://www.bizjournals.com/dallas/stories/2002/12/02/focus3.html on 11 June 2003.

[132] United States Bureau of the Census. http://www.census.gov.

[133] United States Congress. Aviation and Transportation Security Act. *Public Law 107-71*, 107^{th} *Congress*, 19 Nov. 2001.

[134] United States Congress, Office of Technology Assessment. Technology Against Terrorism: Structuring security. Technical Report OTA-ISC-511, U. S. Government Printing Office, Washington D. C., Jan. 1992.

[135] United States Department of Homeland Security. Fact Sheet: United States Department of Homeland Security Fiscal Year 2006 Budget Request Includes Seven Percent Increase. 2005. Accessed at http://www.dhs.gov/dhspublic/display?theme=43&content=4337&print=true on 27 Feb. 2005.

[136] United States Department of Homeland Security Office of the Inspector General. Audit of passenger and baggage screening procedures at domestic airports. *Office of Audits*, (OIG-04-37), Sept. 2004.

[137] United States Department of Homeland Security Office of the Inspector General. Transportation Security Administration: Review of the TSA passenger and baggage screening pilot program. *Office of Audits*, (OIG-04-47), Sept. 2004.

[138] United States Department of Homeland Security Office of the Inspector General. Follow-up audit of passenger and baggage screening. *Office of Audits*, (OIG-05-16), Mar. 2005.

[139] United States Department of Homeland Security Transportation Security Administration. Report to Congress on enhanced security measures. 2002. Accessed at http://www.tsa.gov/interweb/assetlibrary/Report_to_Congress_on_Enhanced_Security_Measures.doc.

[140] United States Department of Homeland Security Transportation Security Administration. New notice outlines changes to CAPPS II system. *Press Release*, 31 July 2003. Accessed at http://www.tsa.gov/public/display?theme=8&content=631 on 12 Aug. 2003.

[141] United States Department of Homeland Security Transportation Security Administration. TSA canine teams screen US mail for explosives - Pilot program to expand to airports across the country. *Press Release*, (TSA 03-34), 29 May 2003.

[142] United States Department of Homeland Security Transportation Security Administration. TSA selects Lockheed Martin Management and Data Systems to build TSA passenger pre-screening system. *Press Release*, (TSA 15-03), 28 Feb. 2003.

[143] United States Department of Homeland Security Transportation Security Administration. TSA's CAPPS II gives equal weight to privacy security. *Press Release*, (TSA 03-04), 11 Mar. 2003.

[144] United States Department of Homeland Security Transportation Security Administration. TSA to test new passenger pre-screening system. *Press Release*, 26 Aug. 2004.

[145] United States Department of Justice. Terrorism in the United States, 1999: 30 years of terrorism. *Report*, 1999.

[146] United States General Accounting Office. AVIATION SECURITY: Vulnerabilities still exist in the aviation security system. *Testimony of Gerald L. Dillingham Before the Committee on Commerce, Science and Transportation, Subcommittee on Aviation; United States Senate*, (GAO/T-RCED/AIMD-00-142), 6 Apr. 2000.

[147] United States General Accounting Office. AVIATION SECURITY: Vulnerabilities and potential improvements for the air cargo system. *Report to Congressional Requesters*, (GAO-03-344), Dec. 2002.

[148] United States Government Accountability Office. AVIATION SECURITY: Secure Flight development and testing under way, but risks should be managed as system is further developed. *Report to Congressional Committees*, (GAO-05-356), Mar. 2005.

[149] United States House of Representatives. Profiling for public safety: Rational or racist? *Hearing before the Subcommittee on Aviation, Committee on Transportation and Infrastructure*, 27 Feb. 2002. Accessed at http://commdocs.house.gov/committees/Trans/hpw107_64.000/hpw107-64_0.htm on 16 Sept. 2003.

[150] United States Office of National Drug Control Policy. Measuring the deterrent effect of enforcement operations on drug smuggling, 1991-1999. Aug. 2001.

[151] United States State Department. Significant terrorist incidents 1961-2001: A brief chronology. Accessed at http://www.state.gov/r/pa/ho/pubs/fs/5902pf.htm.

[152] United States State Department. Political violence against Americans. *Annual Reports*, 1997, 1998, 1999, 2000, 2001.

[153] United States State Department. Patterns of global terrorism, 1999. *Report*, 1999.

[154] University of Maryland - College Park. EPA national time use survey, 1994-1995. *Scientific Research on the Internet.* Available at http://www.webuse.umd.edu and http://www.popcenter.umd.edu/cgi-bin/hsda?harcsda+time.

[155] J. S. Vandergraft. A fluid flow model of networks of queues. *Management Science*, 29(10):1198–1208, Oct. 1983.

[156] J. L. Virta, S. H. Jacobson, and J. E. Kobza. Outgoing selectee rates at hub airports. *Reliability Engineering and System Safety*, 76(2):155–165, 2002.

[157] J. L. Virta, S. H. Jacobson, and J. E. Kobza. Analyzing the cost of screening selectee and non-selectee baggage. *Risk Analysis*, 23(5):897–908, Oct. 2003.

[158] M. L. Wald. A Nation Challenged: AIRLINE SECURITY; Official says he'll miss screening goal. *The New York Times*, page B9, 28 Nov. 2001.

[159] M. L. Wald. Tough issues on baggage screening remain. *The New York Times*, page A15, 5 Nov. 2002.

[160] M. L. Wald. Travel Advisory: Correspondent's Report; A plan for screening at airports in dropped. *The New York Times online*, 1 Aug. 2004. Accessed at http://www.nytimes.com on 1 Aug. 2004.

[161] W. L. Waugh, Jr. Securing mass transit: A challenge for homeland security. *Review of Policy Research*, 21(3):307–316, May 2004.

[162] White House Commission on Aviation Safety and Security. Final Report. 12 Feb. 1997.

[163] W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212, 1999.

[164] M. Williams, J. M. Johnston, L. P. Waggoner, M. Cicoria, S. F. Hallowell, and J. A. Petrousky. Canine substance detection: Operational capabilities. In *Office of National Drug Control Policy Technology Symposium*, 1999.

[165] T. Zeller. Cheap and lethal, it fits in a golf bag. *The New York Times online*, 26 Oct. 2003. Accessed at http://www.nytimes.com on 29 Oct. 2003.