To the Graduate Council:

I am submitting herewith a thesis written by Linda Jean Dupont entitled "An Expert System for International Mail Loss Investigation." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.

Michael G Thomason

Michael Thomason, Major Professor

We have read this thesis
and recommend its acceptance:

Brad Vander Zanden

Accepted for the Council:

Lee Minkel

Associate Vice Chancellor
and Dean of the Graduate School

An Expert System for International Mail Loss Investigation

A Thesis

Presented for the

Master of Science Degree

The University of Tennessee, Knoxville

Linda Jean Dupont

August 1997

Acknowledgments

- First I would like to thank my advisor, Dr. Michael Thomason, for his assistance with this project.
- Thank you also to Dr.'s Straight and Vander Zanden for being on my committee.
- Many thanks to Ryan Carroll for his patience and help while I finished this.
- To Ethel Wittenberg and Wallace Mayo for their encouragement and faith that I would make it.
- To Dennis Hitzing , "I'm sure your as glad as me that it's finally over!"
- And thanks to anyone else who deserves it but I may have forgotten.

Abstract

Postal services around the world experience the loss of international mail due to pilferage which occurs during the mail's routing. The routing of international mail is complex with the mail being handled by at least two postal administrations and usually at least one commercial airline contracted to transport mail across international borders. With all this handling it is often difficult to locate the source(s) of pilferage along the mail's route.

In this thesis we describe the development of, and present, one possible approach that can assist postal inspectors to locate the segment(s) in the routing of international mail where the pilferage may be occurring. Our approach uses simple statistical methods to systematically analyze the routing of international mail. The analysis begins at the location where the loss was detected and traces back along the mail's route until a likely source is detected or the dispatch location is reached.

Contents

1 Introduction **1**

2 Postal Service **3**

2.1 Historical Overview 3

2.2 International Mail 5

2.3 Mail Loss Investigation 5

3 Developing An Expert System **8**

3.1 Expert System Development 8

3.2 Consultation with an Expert 12

3.3 Mean and Standard Deviation 16

3.4 Basic Algorithm Developed 17

3.5 The Prototype System 19

3.6 Testing of Prototype 20

4 From Prototype to Full System **22**

4.1 System Adaptations 22

4.1.1 Case 1: Small Number of Possible Pilferage Sources 23

4.1.2 Case 2: Evaluation Sensitivity 23

4.2 Graphical User Interface 25

4.2.1 Language Conversion 25

4.3 Features 26

4.3.1 Display Options 27

4.3.2 Analysis Options 29

4.3.3 Other Features 31

5 Conclusion **33**

5.1 Current Status 34

5.2 The Future 35

Bibliography **36**

Appendix **39**

A Meeting Minutes **40**

A.1 Topics of Discussion: Available resources

Date: January 23, 1996 40

A.2 Topics of Discussion: Graphical Representation

Date: January 29, 1996 41

A.3 Topics of Discussion: Project Development

Date: February 22, 1996 42

A.4 Topics of Discussion: Data Format and Analysis

Methods

Date: August 02, 1996 44

A.5 Topics of Discussion: Error Checking and Algorithm

Date: August 19, 1996 48

B Mr. Hitzing's Example **50**

C Data Input Format **52**

Vita **53**

List of Tables

3.1 Expert System Steps 9

3.2 Meeting Dates and Topics 12

3.3 Basic Prototype Algorithm 19

4.1 Receptacle Event Codes 28

C.1 Data Input Format 52

List of Figures

3.1 Graphical Representation of Expert's Concept 14

3.2 Graphical Representation 15

3.3 Graphical Representation used in Prototype 18

4.1 Chart of System's Operational Flow 26

Chapter 1

Introduction

Once mail has been entrusted to a postal service, we expect it to be delivered safely to its destination. Most of the time this is exactly what occurs, but unfortunately it is not always the case. Sometimes it arrives a little worse for wear, other times it does not arrive at all. The reasons for this are numerous. For example, the delivery address could be inaccurate, unreadable, or missing; the parcel could be delivered to the wrong address; it could be lost; or it could be pilfered or stolen.

Postal services have become part of daily life all over the world. Thus, recognizing the international importance of the reliable transport of mail, the Universal Postal Union (UPU) has set up a task force to investigate international mail irregularities. The Universal Postal Union, an organization of postal administrations from around the world, is a specialized institution of the United Nations that regulates postal services. Since postal security is a high priority of the UPU, it has

created the Postal Security Action Group (PSAG). One of the PSAG's missions is to investigate international mail loss with the view of enhancing the quality and reliability of postal service [10]. The theft or pilferage of mail are irregularities collectively referred to as mail loss. Mail loss is the type of mail irregularity on which this thesis focuses.

Mr. Dennis Hitzing, a specialist-consultant to the PSAG and inspector in the United States Postal Service, has suggested that systematic analysis of mail loss patterns may aid in the identification of the source of these irregularities. The goal of this thesis is to investigate and develop an initial version of an expert system that assists in this analysis.

Chapter 2 provides a brief history of postal service and an overview of international mail and introduces current mail loss investigation techniques. In chapter 3, the development of an expert system that supports mail irregularity investigation and the steps taken to develop such a system are discussed. This is accompanied by the algorithm, and an explanation of the algorithm, that was developed for use in the prototype system. Chapter 4 describes the progression from prototype to fully operational system, the features incorporated into the system, and the graphical interface developed. Finally chapter 5 discusses the current status of the system and possible future work on the project.

Chapter 2

Postal Service

A postal service is responsible not only for the collection and distribution of mail but also for investigating and correcting mail irregularities so that mail is delivered safely to its destination.

This chapter describes the evolution of postal service from mail transported by messengers on foot to the complex international mail service we have today. The current method of mail delivery and of mail loss investigation are also discussed.

2.1 Historical Overview

There are historical references to postal systems as far back as the second millennium BC in Egypt and the first millennium BC in China. Typically such systems were originally created by governments to maintain correspondence between their political hubs and their military in distant regions [1].

In the early days, mail was delivered by messengers on foot, on horseback, by mule, or in horse-drawn carriages. The number of persons who handled the mail and the volume of mail being handled were relatively small by today's standard. Thus when mail was pilfered or missing, the list of suspects was short.

As transportation improved in the 19th century, the steamship and railroad were utilized in the delivery of mail. The 20th century brought about the use of automobiles, trucks, and airplanes. As populations increased and dispersed, the volume of mail increased. As a result the number of people involved in the handling of mail increased and the routes by which it was delivered became more segmented and complex.

In modern times, mail is brought to post offices by individuals, businesses, or postal employees. At the post office, letters are separated from parcels and divided by type. Airmail, first class, second class, and so on are all separated. Next the mail is postmarked and sorted according to its destination. Once sorted, there may be several modes of transport involved in delivering the mail to its final destination. All this leads to an extensive system of handling and routing which complicates the task of locating the source of mail irregularities when they arise.

2.2 International Mail

International mail also has the complication that, at some point in its handling, the mail will most likely leave the possession of the postal service and be entrusted to a commercial airline travelling across international boundries. In all cases, at least two postal services are also involved, that of the delivering country and that of the receiving country.

Today, the United States Postal Service has contract space on approximately 15,000 of the 56,000 worldwide commercial airline flights daily [9]. Worldwide, airlines transport over 10 million kilograms of mail on an average day [2]. The large volume of mail, the vast number of different flights involved in delivering mail internationally, and the involvement of multiple postal services has made investigation of mail loss a difficult task. Since an irregularity or loss may occur at any of several stops along the mail's routing, a postal inspector has quite a chore in identifying the source of losses.

The next section looks in more depth at the techniques currently used by postal inspectors in the investigation of mail loss.

2.3 Mail Loss Investigation

Locating the source of the irregularity, be it damaged mail, pilfered mail, or lost mail, begins with the filing of a report. The report contains information on when

the irregularity was discovered, what the irregularity was, and the route by which the mail was transported. Originally the reports were completed by hand and then entered manually into a database, a process which has led to incomplete and inaccurate data in database. The database was then used to produce lists of reports which all originate at the same location. This list would then be passed to a postal investigator to locate the source or sources causing the generation of the reports. However there were no tools available that could help pinpoint, or even narrow down the search as to cause of the irregularities or losses.

Today, technological advances have made it easier to submit accurate, complete reports through the use of bar-code scanners and electronic submission methods; nonetheless, the mail services' databases still typically provide only lists of reports for a postal inspector. A postal inspector must still search through large volumes of data in the hope of narrowing the search to the possible source of mail losses or irregularities. For example, if a large number of reports is being generated from a specific location, a postal inspector must manually sort through all the reports generated. He will go through the reports trying to identify recurring patterns and try to pinpoint the source of the losses or irregularities causing the reports' generation. Next an inspector will go to a suspected source of the losses to try to locate the problem and correct the situation.

In the next chapter the goal of utilizing current technologies to assist in the analysis of mail loss, the restraints imposed on the analysis, and the process by

which such a system was developed are discussed.

Chapter 3

Developing An Expert System

This chapter outlines the development of a prototype expert system to assist postal inspectors in identifying possible problem sites in the transport of international mail. An expert system is a computer system that can help solve complex, real world problems in specific scientific, engineering, or medical specialties. Such systems are characterized by their use of facts and procedures, obtained from human experts, that have proved useful for solving typical problems in their domain [4].

3.1 Expert System Development

A very basic outline of the steps to developing an expert system are listed in Table 3.1 [8, 4].

The problem description is, initially, an expert's written description of the problem to be solved. The problem to be solved can also be referred to as the

Table 3.1: Developmental Steps for Expert System Design

Stage	Task
step 1	Problem Description
step 2	Collaborative Refinement
step 3	Knowledge Acquisition
	a: Interview Expert(s)
	b: Check Documentation
step 4	Prototype System Specifications
step 5	Prototype Development
step 6	Testing
step 7	Evaluation
step 8	Final System Specifications
step 9	Final System Development
step 10	Technology Transfer

domain task. A collaborative refinement of the description is a clarifying of the problem description through discussions between the expert and the system developer. The goal of the refinement is to make the description more accurate, understandable and straightforward to assist all involved in understanding the problem description thoroughly.

Knowledge acquisition is one of the more difficult tasks in the development of an expert system. It involves interviewing experts to investigate the domain task in depth and to glean knowledge of their procedures used in solving the problem. The latter part is not as easy to accomplish as it may at first appear. It is common for experts' explanations to be based on assumptions and/or knowledge that is obvious to them but not at all obvious to a layman.

Knowledge acquisition also involves the investigation of any documentation

available that may be relevant to the domain task. The documentation can be in user's manuals, office procedure manuals, or similar material.

Only after the domain task has been investigated and knowledge acquired can the task of designing the system begin. The first thing required is system specifications. System specifications consist of what the input will be, what the expected output is, and any additional data that will be used by the system.

Initially only a prototype of the system is built. The prototype generally consist only of the "bare bones" required to test and evaluate the system. Usually the prototype will not contain all the features and functions found in the final version of the expert system. Testing involves checking of the prototype to see if it performs as the system developer expected it to. System evaluation is the checking of the system to see if it performs as the expert expected it to.

These initial seven steps comprise the "prototyping" cycle, and any one step's development may cause the return to a previous step to make adjustments.

Once a satisfactory prototype has been developed, a final set of system specifications will be developed that will contain all the specifications of a final version of the expert system. The final system specifications will containing a listing of all the functions and features the expert feels the system should provide. The final version of the system is then implemented with as many of the features as possible and with an appropriate user interface to make the system accessible.

The last step is in the transfer of the technology from the system developer

and expert to others who may use the system by providing documentation, system guides and user's manuals.

Not listed in the table is "communication" or "development of a common language". This must be done early on and refined throughout the entire development process. It is common initially for the field experts and the system developer to have a period of education which involves learning each other's vocabulary. Experts in most fields develop a vocabulary that is pertinent/relevant to their specialty and that people outside of the specialty may not fully understand. The lack of understanding can be due to several factors such as implied associations with certain terms, instance specific meanings or uncommon meanings to familiar words or phrases.

Some difficulties often encountered in developing an expert system lie in knowledge acquisition. Being human, experts tend to over-simplify the problem, may leave out relevant facts, and could express the problem in such a way as to mislead or confuse the developer as to exactly what the problem is. Over time, experts can develop a familiarity with and background knowledge of the problem. They will use this subconsciously in the problem solving process. Another difficulty that may be encountered is inaccuracy and incompleteness of available documentation.

Table 3.2: Dates and Topics of Meetings held with Mr. Hitzing

Date	Topics of Discussion
January 23, 1996	Available Resources
January 29, 1996	Graphical Representation
February 22, 1996	Project Development
August 02, 1996	Data Format and Analysis Methods
August 19, 1996	Error Checking and Algorithm

3.2 Consultation with an Expert

Consultation with Mr. Hitzing, a US Postal Inspector experienced in mail loss investigation, was an extensive portion of the initial developmental process. Table 3.2 contains a listing of some of the meetings held with him.

The development, as can be seen from the meeting minutes found in their entirety in Appendix A, initially consisted of becoming familiar with the problem of international mail loss and developing a common vocabulary. This period was essential to developing a process to simulate some aspects of Mr. Hitzing's analysis and therefore provide him with a work aid.

During this developmental process, several possible analysis procedures were evaluated and rejected because they would have required unavailable, or difficult to obtain, data or the procedure, upon closer examination, did not fill the ultimate requirements of the system. For example, it was suggested that mail volume be incorporated into the analysis process to calculate the percentage of the total mail being transported along a route experiencing pilferage. Mr. Hitzing rejected this

possibility because his goal was to target the locations where most pilferage was occurring independent of the volume of mail transported via that route. To get a better understanding of how Mr. Hitzing analyzed mail loss, he was asked to describe his analysis procedure.

When it came to conceptually explaining his analysis procedure, Mr. Hitzing had difficulty. There were several reasons for this. First he had difficulty explaining the procedure he used in the analysis of mail loss because the process had become routine for him. Second, it was not always explicitly clear what personal knowledge he incorporated into the analysis process. Appendix B is the process he used expressed in his own words.

His analysis began with focusing on a group of reports and trying to locate patterns among the reports. Mr. Hitzing categorized reports according to where the pilferage was reported and the route over which the pilfered mail had been transported. This is illustrated by the graph in Figure 3.1 which Mr. Hitzing produced to visually depict his analysis process. The nodes' are labelled with an airpot's code and the arcs are labelled with airline flights. In the graph he used nodes to represent groups of reports that had all passed through the same airport.

Once Mr. Hitzing had produced a graphical representation, he would start at the site where the pilferage reports were generated and trace back along the mail's routing searching for patterns of handling common to a majority of the reports. Some of the elements of handling he would investigate were a common

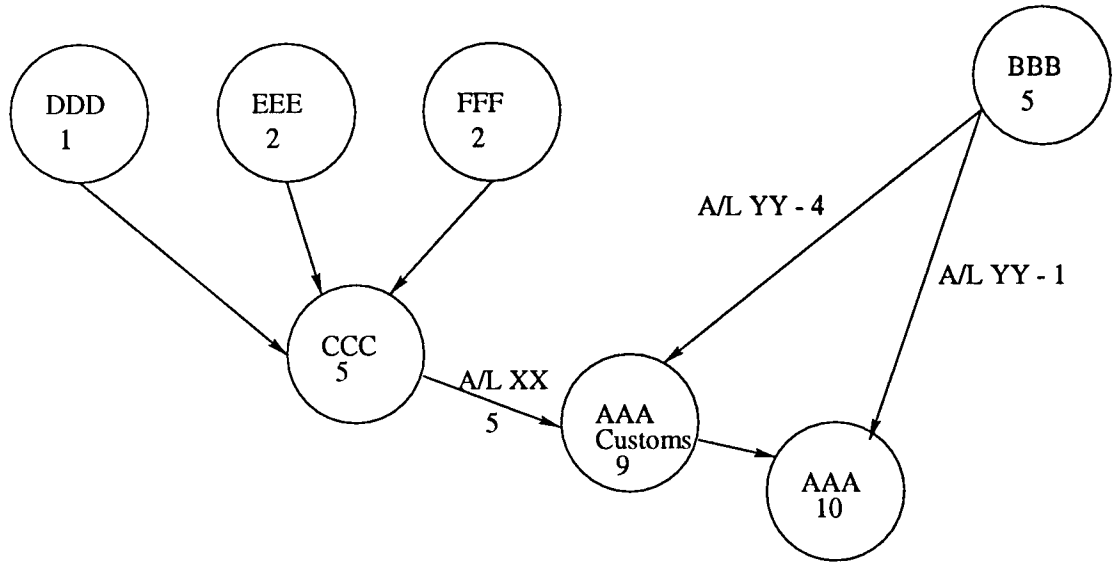


Figure 3.1: A reproduction of the graphical representation envisioned by Mr. Hitzing. The arcs are labelled with airline flights and the nodes with airport codes.

airline transporting a majority of the mail and a common airport that the mail was transported through. If a pattern was found, he would recommend that an investigation target the location of his suspicions.

After Mr. Hitzing had produced his initial description of the analysis process, numerous meetings spanning several months were spent collaboratively refining it. Part of this involved improving the clarity of the domain task description. Creating a mental picture and then a graphical one helped in visualizing and defining the problem. Initially, we each developed our own images of the problem and through discussions combined them to form a mutually agreeable image and problem description. The rough graphical representation that was developed is depicted in Figure 3.2. This representation did not differ greatly from Mr. Hitz-

ing's initial image. It can be seen that the structure evolved from his image is a basic hierarchical tree representation where a single flight is represented by three levels of the tree in which the main level is the destination airport for the flight, the next is the airline or carrier, and the final level is the specific flight number.

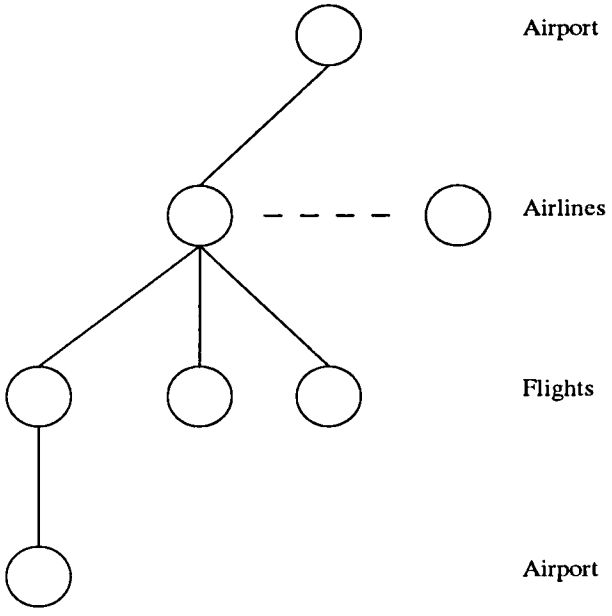


Figure 3.2: The graphical representation developed through discussions with Mr. Hitzing.

Once a graphical representation was developed, the next step was to develop an algorithm for data analysis. The graphical representation helped in understanding the domain task and in investigating statistical methods that might simulate or support some aspects of Mr. Hitzing's logical analysis procedure. In viewing the graph, it was noticed that Mr. Hitzing tended to follow arcs that had an uncommonly large number of reports associated with them. This led us to investigate the use of mean and standard deviation in the analysis process to flag the arcs

that should be marked as possible sources of pilferage.

3.3 Mean and Standard Deviation

When looking for an analysis procedure, we wanted to use well established statistical methods. Mean and standard deviation are fundamental in statistical analysis. Mean or, more specifically, arithmetic mean is a statistic that describes the central tendency of a data set. However, mean is easily skewed and alone it would not be sufficiently accurate for our application. Therefore, we incorporated standard deviation into the analysis because we were interested in how the values of the data set vary from one another. Standard deviation is a measure of the dispersion or variation in a data set.

Under certain assumptions, standard deviation reflects the probability that an event is occurring randomly. Conversely, it is a comparative measure of the probability that an event is occurring non-randomly, or in this case that pilferage is a deliberate recurring act. To provide a comparative standard, the assumption we are making is that the occurrences of pilferage would exhibit maximum entropy if they were random acts. Maximum entropy, which has value log n for n discrete options, corresponds to maximum uncertainty associated with a system. In our case, that would correspond to all possible sources of pilferage appearing to be equally likely [7].

In simple test cases, mean and standard deviation appeared to simulate sufficiently the process used by Mr. Hitzing. Therefore, we decided to proceed to prototype using an algorithm based on mean and standard deviation.

The algorithm developed for use in the prototype is presented and explained in the following section.

3.4 Basic Algorithm Developed

To make the implementation of a prototype as straightforward as possible it was decided to simplify the graph structure to contain only one type of node instead of the three node types shown in Figure 3.2. The nodes would represent airports and the arcs would represent any flight between two given airports. It should be noted that no information is lost due to the simplified graphical representation used in the development of a prototype. Information that had been explicitly represented in the original representation is now implicitly associated with the arcs in the simplified representation.

With this simplified graphical representation, shown in Figure 3.3, values would be assigned to arcs. The value assigned to a particular arc would correspond to the number of reports associated with the route segment the arc represents, this will be referred to as the arc's value. Once all arc values have been assigned, the mean of all values assigned to the arcs leading into a single node would be

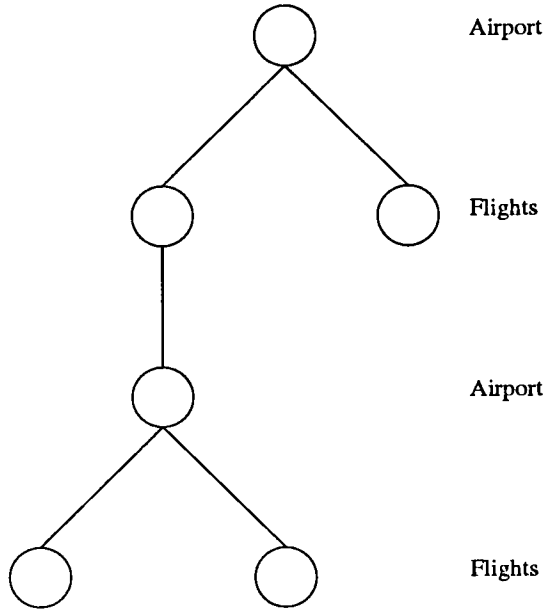


Figure 3.3: The simplified graphical representation used for development of the prototype.

calculated. Next, the number of standard deviations each arc value is away from the mean would be calculated.

The number of standard deviations an arc value is away from the mean is taken to reflect the arc's likelihood of being a source of pilferage. If an arc's value is zero or fewer standard deviations from the mean, it has fewer than the average number of reports associated with it and will not be investigated further as it is unlikely to be a source of pilferage. If an arc is a positive number of standard deviations from the mean, it has a larger than average number of reports associated with it, and therefore may be a source of pilferage.

The algorithm used to develop the prototype is presented in Table 3.3.

Table 3.3: For each probable recipient of pilfered mail, the following is the basic algorithm developed to identify the most likely source(s) of the pilfered mail.

Step	Procedure
1	Calculate the mean and standard deviation of the report count of each of the route segments leading into the site currently being analyzed. The report count of a route segment is defined to be the number of reports currently under consideration that are associated with that segment.
2	Compute the number of standard deviations between the mean of the report count and each route segments individual report count.
3	Mark all route segments whose report count deviates from the mean by more than the currently specified threshold as a possible recipient of pilfered mail.
4	For each marked route segment remove from consideration all reports not associated with that route. Return to step 1.

3.5 The Prototype System

Prior to the development of the prototype, Mr. Hitzing and I met several times in order to finalize the specification of the input and output. To minimize the impact on an already established system of pilferage reporting, it was decided that the program's input should be derived almost directly from the standard reporting form currently in use by Postal Administrations around the world. However, Mr. Hitzing did request that the Technical Advisory Board of the Universal Postal Union accept some suggested changes to the standard reporting form. For example, he requested that the mail's routing data become a required field because this data is necessary for proper analysis of reports by the program. Some of the other changes suggested were the addition of new receptacle event codes to better

classify the type of damage or pilferage detected, and the restructuring of the format to one that is more computer readable. The final input format settled upon is provided as Appendix ??.

Our main goal in developing a prototype was to be able to test the proposed algorithm. Thus, we were more concerned with content than with appearance. The output is the only window into the operation of the system and is all the user has to evaluate the system's performance. The output, for this reason, needs to be comprehensive and readily understandable. Taking this into account, the format of the output very much resembled debugging messages, reporting the full state of the program at critical points in the evaluation of the input data.

For ease of development the prototype was initially implemented in the language C as a command line program. A graphical user interface is desired in the final system but was not necessary for testing of the prototype; therefore, it was left for later development.

3.6 Testing of Prototype

One of the first things discovered upon loading test data was that the input did not necessarily contain all specified required fields. Due to this fact, before testing could proceed, an input filter needed to be incorporated into the prototype. The filter would be used to check input files and reject any report not containing all

the required fields.

When testing recommenced, the program ran much as expected except in certain cases. The first exception was that it did not handle the evaluation of possible sources of pilferage well when there were only a small number being investigated. Another case was that it was not sufficiently sensitive in marking the possible sources of pilferage.

In the next chapter, the methods developed to handle these cases more effectively are presented along with changes and improvements made to the program in the transition from prototype to fully operational system.

Chapter 4

From Prototype to Full System

This chapter describes the progression from prototype to fully operational system. We discuss the adaptations that were necessary to handle system shortcomings detected during testing and describe the features of the fully operational system. The design of the interface is also discussed.

4.1 System Adaptations

There were two main cases discovered during testing of the prototype where the system did not perform as well as expected. The steps taken to correct the performance of the algorithm in each of these cases are outlined in the following two subsections.

4.1.1 Case 1: Small Number of Possible Pilferage Sources

The performance of the prototype algorithm seriously deteriorated on groups of one or two. These were for different reasons but the affect was the same, the prototype algorithm always stopped evaluation. For a group of one, evaluation ends because when there is only one possible source of pilferage it is obviously a suspect route segment. While this behavior may at first appear correct, it was not desirable. We wanted the evaluation to continue because the true source of the pilferage may be prior to the current segment in the mail's routing. This was overcome by always continuing evaluation on groups of one.

For groups of two, evaluation ended whenever the standard deviation threshold was set above one standard deviation because each possible source is always one standard deviation from the mean. In certain instances of this case, one possible source of pilferage was obviously suspect compared to the other because it had significantly more reports associated with it. The ending of evaluation was overcome in this case by investigating a source if the number of reports associated with it was more than twice the number of reports associated with the other source.

4.1.2 Case 2: Evaluation Sensitivity

In the case where the system was not always sensitive enough to mark appropriate sources of possible pilferage, it appeared that a more sensitive statistical analysis procedure may be needed.

We were interested in a statistical method that would improve the system's performance without requiring major alterations. We decided to use the Kullback, or discrimination, function which can be used to compare probability distributions [5, 6]. If $P = \{p_1, p_2, \ldots\}$ and $Q = \{q_1, q_2, \ldots\}$ are both distributions defined on the same sample space then the discrimination with respect to P is

$$L(P,Q) = \sum_j p_j \log \frac{p_j}{q_j}.$$

The Kullback function is also referred to as a relative entropy measure since $L(P,Q) = 0$ if and only if $p_j = q_j$ for all j. This is the sense in which we are interested in the function. We view the possible sources of pilferage as distributions. P reflects a distribution exhibiting maximum entropy and Q reflects the actual distribution of the sources. Thus $L(P,Q)$ is the relative entropy of the actual distribution, Q, with respect to maximum entropy, P. If the relative entropy measure is zero then we know there is maximum entropy; otherwise, the larger the measure, the greater the discrimination of the distribution. In other words, the greater the relative entropy measure, the greater the likelihood that at least one of the possible sources is suspicious.

The prototype algorithm was adjusted by first using the Kullback function to determine whether or not any source is suspicious and should therefore be investigated. Only when the Kullback function indicates suspicion is the mean

and standard deviation for the distribution calculated. If, when using the original evaluation method, no source appears suspicious, then the source with the greatest number of reports associated with it is investigated.

4.2 Graphical User Interface

A major feature desired in the fully operational system but not incorporated into the prototype was a graphical user interface. The aim of incorporating a graphical user interface would be to increase the ease of use of the system. It is possible that some of the eventual users of the system may have very little computer experience.

4.2.1 Language Conversion

The prototype system was developed in the language C but it was decided to convert to the Java language for the next stage of system development. There were several reasons for this decision. First, Java allows quick development of graphical user interfaces and is portable to IBM compatible personal computers, Macintosh computers, and many Unix-based systems. Second, conversion of C to Java would not require extensive alterations to the existing code since Java shares much of the same syntax as C.

4.3 Features

It was decided that the fully operational system would be flow driven. The chart in Figure 4.1 depicts the system's operational flow.

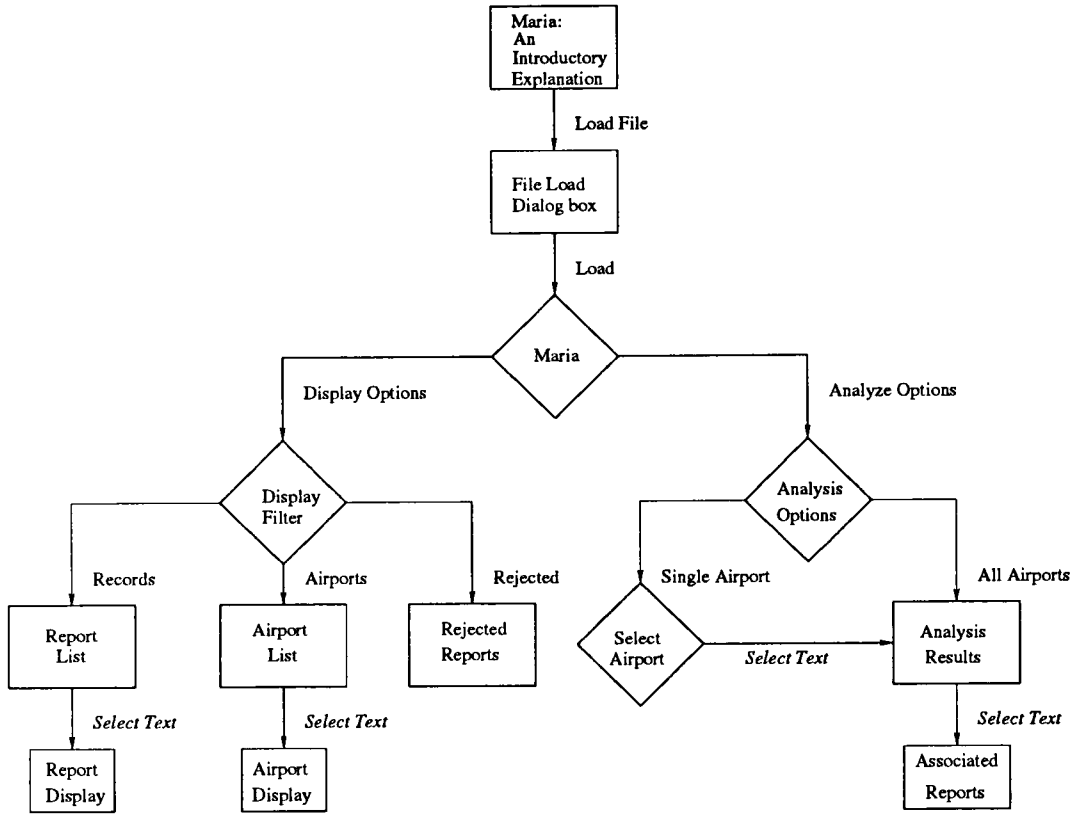


Figure 4.1: This chart depicts the system's operational flow. Arcs labeled in italics reflect the selection of text, others reflect button selection.

The first thing that happens after starting the program is that an introductory explanation window appears. This window contains a brief explanation of the program's purpose. It also contains a "Load File" button which upon selection closes the current display and invokes a built-in Java file dialog box. This file

dialog box can be configured to handle the loading or saving of files. In this instance it was configured to handle the loading of a file into the system. The file dialog box visually depicts the contents of the current working directory and allows the user to select a file for loading. After this initial load, additional files can be loaded by selecting a menu option in the main window.

Once a file has been selected for loading, the system provides some transparent features for the user. The first is that it can handle empty or improperly formatted files. Second, it filters the reports contained in a file to ensure that only reports containing all required fields are loaded into the system. This second feature was incorporated during the testing of the prototype system when it was discovered that it could not be assumed that all reports would contain all the required fields.

After loading an initial file, the system displays its main window. This window provides a brief description of the main options available to the user.

At this stage there are two main options, "Display Options" and "Analyze Options", that are selectable by the user.

4.3.1 Display Options

The main purpose of this option is to provide the user with a means of displaying filtered listings of the data that has been loaded into the system.

Upon selecting "Display Options" the display filter window is shown. This window consists of two main sections: a filter and the display options. The filter

Table 4.1: Receptacle event codes. Column two indicates which are selectable only in the analysis option display window.

Code	Selectable	Description
20		Received here in good condition: cancel any code 31 entry
21	x	Damaged by weather or improper handling: NO theft evident
22	x	Received cut, torn, or broken: enclosed mail NOT intact
23	x	Seal tampered or missing: enclosed mail NOT intact
24	x	NO external tampering detected: enclosed mail NOT intact
25		Received here with unreadable bar code
26		Pilfered at this location: found evidence of local theft
27		Found unprotected or abandoned at this location
28		Carrier failed to make transfer intended at this location
29	x	Misdelivered to this location
30	x	Received here without receptacle label: copy of substitute sent
31	x	Did not arrive here: manual report of missing receptacle

and display options are described in the following two subsections.

The Filter

Currently filtering by receptacle event code and/or dispatch date of the inputted reports is supported. The receptacle event codes currently defined are listed in Table 4.1. The default setting of the filter has all receptacle event codes selected and the dispatch date selected range begins on January 1, 1970 and ends on the present date.

The Display Options

There are three display options: "Report", "Airport", or "Rejected". The "Rejected" option refers to a listing of all the reports that were rejected by the system during loading. This display is not filterable due to the fact that some of the fields required by the filter may not have been supplied. The "Report" and "Airport" options both filter the reports they include in their listings. The "Report" listing contains all the reports that met the selected filter requirements. The "Airport" listing contains all the reports that met the selected filter requirements grouped by the airport at which the report was generated. In both the "Report" and "Airport" displays, if the user selects a particular report or airport, a window displaying more detailed data on the selected item will be displayed.

4.3.2 Analysis Options

The main purpose of this option is to invoke an analysis of a single airport or of all the airports. The options window displayed upon selecting "Analyze Options" consists of three main sections: a report filter, variable thresholds, and the analysis options. The report filter is similar to the "Display Options" filter except that not all receptacle event codes are selectable. This is because certain receptacle event coded reports are not relevant to international mail loss investigation. An example of an irrelevantly coded report is a report with receptacle event code 20, "received here in good condition: cancel any code 30 entry". All the relevant

receptacle event codes are indicated as such in Table ??. The options threshold setter and analysis options are described in the following two subsections.

Analysis Thresholds

Variable thresholds are variables that may be set within a predefined range by the user and are used in the analysis of airports. The *minimum number of reports* variable refers to the minimum number of reports required to be associated with a possible source of pilferage for that source to be investigated further. The default setting for this threshold is 6. The *standard deviation* variable is the threshold level above which a possible source of pilferage is considered suspicious. This variable's default value is 2. The *Kullback* variable is the relative entropy threshold calculated for a group of possible sources of pilferage currently under investigation. The default value for this threshold is 0.3 and was obtained through testing and evaluating after its incorporation into the system.

The user, after setting the thresholds and filter options to desired levels or values, can chose to either analyze a specific airport or analyze all airports.

Analysis Options

There are two analysis options: "Single Airport" or "All Airport". The "Single Airport" option displays a listing of all the airports that are available for analysis. Once the user selects the airport he desires to analyze, the analysis is performed

and the analysis result window displays the results. If the user selects the "All Airport" option, then an analysis of all the airports is performed and the analysis result window is displayed.

The format used to display results is very similar to the display used in the prototype. After discussing possible result formatting options and reviewing the results provided by the prototype system, we decided to use the basic format used in the prototype system. The reason for this decision was that the display of results in the prototype system provided a full trace of the investigation from the airport where the report was filed back to the suspicious route segment.

As in the "Report" and "Airport" listing mentioned above, the user can select any displayed line in the analysis result window that mentions a group of reports and a window displaying those reports will be displayed.

4.3.3 Other Features

Having a graphical user interface allows many of the system's options to be set or selected using only a mouse. For example, variable thresholds are displayed as scrollbars, receptacle event codes can be selected or deselected by clicking a checkbox, and dates are set by entering the date into the appropriate text field. Other features provided by the graphical user interface are buttons to select current available window options and file menus which allow the saving of the contents of a window or the closing of a window. Most display windows in the

system contain a file menu that has a save feature. This feature allows the content of the window to be saved to a file.

A print option was not incorporated into the system because print is not supported by any of the standard Java packages. Not incorporating a print option allowed us to keep the system portable. The user always has the option of saving the file and then printing it using a system dependent utility.

Chapter 5

Conclusion

The purpose of this project was to develop an initial version of an expert system capable of assisting Mr. Hitzing, and other Postal inspectors, in the investigation of international mail loss. The program uses statistical methods to systematically analyze reports and provide a listing of possible sources of pilferage. The prototype was tested using contrived simple test cases on which it appeared to perform satisfactorily except in a few specific cases. Alterations were incorporated into the algorithm to handle these cases in the fully operational version of the system.

Once the fully operation system was running, it too was tested on contrived data in order to use scenarios where the source or sources of the pilferage were evident to the human expert. Once satisfied that the system performed well on the contrived cases, we began testing the system using actual reports. Initially, we limited these cases to a small number of reports relating to already suspect

routes. The system was eventually run on a file containing more that 3,000 actual reports. Mr. Hitzing's judgement was that the system performed satisfactorily on this large collection of reports.

At this stage it was decided that the system was ready for use and further evaluation (beta testing) by Postal inspectors.

5.1 Current Status

The system is currently installed on Mr. Hitzing's office computer and the computer of Mr. Ron Murphy, the Universal Postal Union Security Secretary at the Universal Postal Union's headquarters in Bern, Switzerland. There has also been interest shown by other postal administrations from such countries as Australia and the United Kingdom to participate in the beta testing of the system.

The Universal Postal Union holds semiannual meetings each spring and fall at its headquarters in Bern, Switzerland. Many different organizations attend this meeting and many make their own presentations. The system was described and demonstrated during the spring of 1997. It was presented on April 22, 1997, to the International Air Transportation Association and on April 29, 1997 to both the Postal Security Action Group and Technical Standards Board of the Universal Postal Union. Also, during the meeting, two years, data was loaded into the system. The results appeared promising and "many folks (were) impressed" [3].

5.2 The Future

At some point in the future, features currently not incorporated in the system may be added. These include a hierarchical representation similar to the one displayed in Figure 3.2 and a print option. There is also a possibility that the program will be incorporated into an already existing United States Postal Service system or into one being developed by the Universal Postal Union for use by postal administrations around the world.

Bibliography

Bibliography

[1] Postal service. *Encyclopedia Americana International Edition*, 22:453–455, 1982.

[2] *Reporting and Analyzing Irregularities in the Air Transport of Mail.* United States Postal Inspection Service, Knoxville, Tennessee, April 1997.

[3] Richard Ascolese. *Email Transmission.* United States Postal employee, April 23, 1997.

[4] A. Barr and E.A. Feigenbaum. *The Handbook of Artificial Intelligence, Volume II.* Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1982.

[5] R.E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, Reading, Massachusetts, 1987.

[6] S. Kullback. *Information Theory and Statistics.* John Wiley and Sons, New York, 1959.

[7] Levin and editors Tribus. *The Maximum Entropy Formalis.* Massachusetts Institute of Technology, Boston, Massachusetts, 1979.

[8] Peter Lucan and Linda van der Gaag. *Principles of Expert Systems.* Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1991.

[9] United States Postal Service. Consumer's Guide to Postal Services and Products. Technical Report 201, USA, July 1996.

[10] Universal Postal Union. *About UPU.* <http://ibis.ib.upu.org/AN/APropos.html>, 1996.

Appendix

Appendix A

Meeting Minutes

A.1 Topics of Discussion: Available resources Date: January 23, 1996

Jan 23, 1996

The Department of Computer Science computer workstations that I have access to are Sun Workstations with a Unix operating system. The computer language I prefer to work in is C.

Will need to know what data is available and its format.

The computer system does have some format conversion capabilities:
dos2unix - converts dos formatted files to unix format

There are others available but further research would be need to ascertain their capabilities.

A.2 Topics of Discussion: Graphical Representation Date: January 29, 1996

29 Jan 96

Def'n:

GRAPH: a diagram (as a series of one or more vertices/nodes and edges/ arcs) that represents the variation of a variable in comparison with that of one or more other variables.

DIRECTED GRAPH (digraph): consists of a set of vertices/nodes together with a set of ordered pairs of elements of vertices/nodes called edges/arcs.

WEIGHTED GRAPH: a graph with value(s) assigned to each edge/arc.

To organize information and data about an airline system a weighted graph representation can be used. We can set up the model by having the airports represented by the nodes/vertices and the flights between airports represented by the arcs/edges of the graph. The arcs/edges will be weighted with numbers extracted from the loss reports and other information regarding the flights.

A.3 Topics of Discussion: Project Development Date: February 22, 1996

Feb 22, 1996

Linda Dupont & Michael Thomason, UTKCS

Goal: To develop a prototype system for International Mail Loss Detection using hardware/software already available in the University of Tennessee Computer Science Department. The system is being developed as part of a Master's Thesis. Therefore, the design must be frozen very soon so that implementation can proceed. The following is a list of features which discussions to date indicate are desirable:

1. statistical summary by:
   a. airline
   b. flight #
   c. airport
      - origin of receptacle
      - subsequent airstop
        (excluding airport where pilferage detected)
      - where pilferage detected/reported

2. purging of:
   a. out-dated data
   b. questionable data

3. interface
   a. convenient data-entry
   b. data/result display

It may not be possible to implement all these features in the prototype because work must be done on a schedule to complete a Master's degree. Furthermore, experience with the prototype is likely to identify changes needed in a more complete system.

The prototype is being designed with the expectation that it will be used as the basis to develop a fully functional system at a later date.

Conceptual Development: Currently visualize a directed graph with the nodes representing airlines and the arcs representing flights and mail loss data.

Hardware: Sun Workstation.

Software: C computer language on a Unix operating system.

Data structures: Contain all the data regarding each mail loss incident in a structure. Have each airline be a link in an ordered linked list with a pointer to the next airline and a pointer to a structure which contains a summary of data pertaining to that airline. This summary of data would contain a list of flights that airline has and a list of mail loss incidents connected to that airline.

A.4 Topics of Discussion: Data Format and Analysis Methods
Date: August 02, 1996

Friday, 02 Aug 1996

Universal Postal Union Project Update
Meeting with Dennis Hitzing on 26 July 1996:

Agenda:
I.Data Organization
A.Format of Input Data
B.Graphing Data Hierarchically
C.Graphing of Routes

II.Data Analysis
Maximum Entropy and Deviation

Meeting Minutes:
I. Data Organization

A.Format of Input Data
It was decided that input data should be formatted as laid out in Appendix A.

B.Graphing Data Hierarchically
A hierarchical approach for structuring the data was discussed and it was decided that the hierarchy should initially consist of three levels (airports, airlines, and flights). Additional levels could at a future date be added to the hierarchy (airline terminals, airport regions, etc.).

C.Graphing of Routes
It was proposed that the routes be traced backwards from the report's source to the pilfered bag's origin. It was agreed that this approach would be beneficial for two reasons. First, looking at the report source locations versus the pilfered bag's origin would reduce the number of initial investigation sites. Second, the back propagation approach explores an alternative to the current forward tracing method and may provide a different perspective of the pilferage problem.

II. Data Analysis

Maximum Entropy and Deviation

entropy - a measure of uncertainty of an entire probability distribution.

maximum entropy - maximum uncertainty, all possibilities are equally likely (i.e., they have the same probability).

deviation - noticeable or marked departure from accepted norms or behavior.

It was noted that in the absence of mail-volume figures pilferage probabilities cannot be estimated as relative frequencies; however, it was noted that the "maximum entropy principle" has several advantages in this case. Maximum entropy is equivalent to the assumption that frequency counts should be evenly distributed across arcs and nodes. Such an assumption is typical in cases where no additional information is available to guide an alternative assignment. In response to this suggestion, Mr.Hitzing indicated that for this investigation the use of probabilities was not desirable. He suggested that since it was the aim of the Postal Inspectors to reduce the amount of pilfered mail, frequency counts would be more appropriate. This approach would more readily pinpoint locations where the greatest volume of mail was being pilfered, thus allowing investigators to target the most active felons.

Meeting Results:

I. Data Organization

A.Format of Input Data

The data will be organized in files in which each line represents an entire report. Each line will be 109 characters long consisting of 20 fields as specified in Appendix A. Missing data must be indicated by placing spaces in appropriate character positions.

B.Graphing Data Hierarchically

The route information will be represented in a hierarchical manner as a tree. The top level nodes will consist of airports, the next level will be the airlines associated with the airports, the final level will be the specific flights associated with each airline. Future versions may allow the incorporation of additional levels

that still maintain consistency of data up and down the tree.

C.Graphing of Routes
Within the graph the routes will be represented by a sequence of links from the report source to the mail bag's origin.

II. Data Analysis

Maximum Entropy and Deviation
In analyzing the graph, we will assume initially for purposes of processing and analysis that pilferage is equally likely at all nodes adjacent to a particular location, i.e. maximum entropy is assumed. When the actual, relative pilferage frequencies differ significantly from the assumed probabilities the site will be marked as a possible recipient of pilfered mail. The route segment where the deviation is the highest will be investigated further as a possible source of the pilfered mail.

It is important to note that the "maximum entropy principle" is widely used in situations where the data or knowledge actually available is insufficient to support a biased, but presumably more accurate, assignment of probabilities and/or evaluation of frequency counts.

Appendix A

Positions	Format	Description
1-6	A6	Origin Exchange Office of the Dispatch
7-12	A6	Destination Exchange Office of the Dispatch
13	A1	Mail Category Code
14-15	A2	Mail Sub-Class Code
16	N1	Last Digit of Year
17-20	N4	Serial Number of the Dispatch
21-23	N3	Receptacle Serial Number within Dispatch
24	N1	Highest Number Receptacle Indicator
25	N1	Registered/Insured Indicator
26-29	N4	Receptacle Weight in Multiples of 0.1kg
30-37	N8	Date Dispatch was Prepared
38-45	N8	Date Consignment made to Carrier
46-53	N8	Date Pilferage was Detected
54-55	N2	Receptacle Event Code
56-61	AN6	Carrier Number 1/Flight Number
62-73	A12	Airstops
74-79	AN6	Carrier Number 2/Flight Number
80-91	A12	Airstops
92-97	AN6	Carrier Number 3/Flight Number
98-109	A12	Airstops

A.5 Topics of Discussion: Error Checking and Algorithm Date: August 19, 1996

Monday, 19 Aug 1996

Universal Postal Union Project Update
Meeting with Dennis Hitzing on 05 Aug 1996:

Agenda:
I. Input Data
Error Checking of Input

II. Data Analysis
Mean and Standard Deviation

Meeting Minutes:

I. Input Data
Error Checking of Input
The issue of error-checking input data was raised. Mr. Hitzing stated that the input data would be checked for errors prior to being processed by this application. The error checking will insure that all required fields contain valid data.

II. Data Analysis
Mean and Standard Deviation
sample mean - a value that is computed by dividing the sum of a set of terms by the number of terms.
sample standard deviation - a measure of the dispersion of a distribution given by the square root of the variance.
variance - a measure of the dispersion of a distribution of a random variable, obtained by taking the expected value of the square of the difference between the random variable and its mean.
expected value - the sum or integral, over all the possible values of a random variable, of the product of the value of the variable, or some given function of it, with the probability of that value

For each probable recipient of pilfered mail, it was proposed that the following be investigated as an algorithm to identify the most likely source(s) of the pilfered mail:

1. Calculate the mean and standard deviation of the report count of each

of the route segments leading into the site currently being analyzed. The report count of a route segment is defined to be the number of reports currently under consideration that are associated with that segment.

2. Compute the number of standard deviations between the mean of the report count and each route segments individual report count.

3. Mark all route segments whose report count deviates from the mean by more than the currently specified threshold as a possible recipient of pilfered mail.

4. For each marked route segment remove from consideration all reports not associated with that segment. Return to step 1.

Meeting Results:

I. Input Data

Error Checking of Input

The input data will be checked to ensure all required fields contain valid data by the transit loss database prior to processing by this application.

II. Data Analysis

Mean and Standard Deviation

The approach outlined above is the one currently under development. Adjustments may occur during development and testing.

Appendix B

Mr. Hitzing's Example

February 27, 1996

Linda:

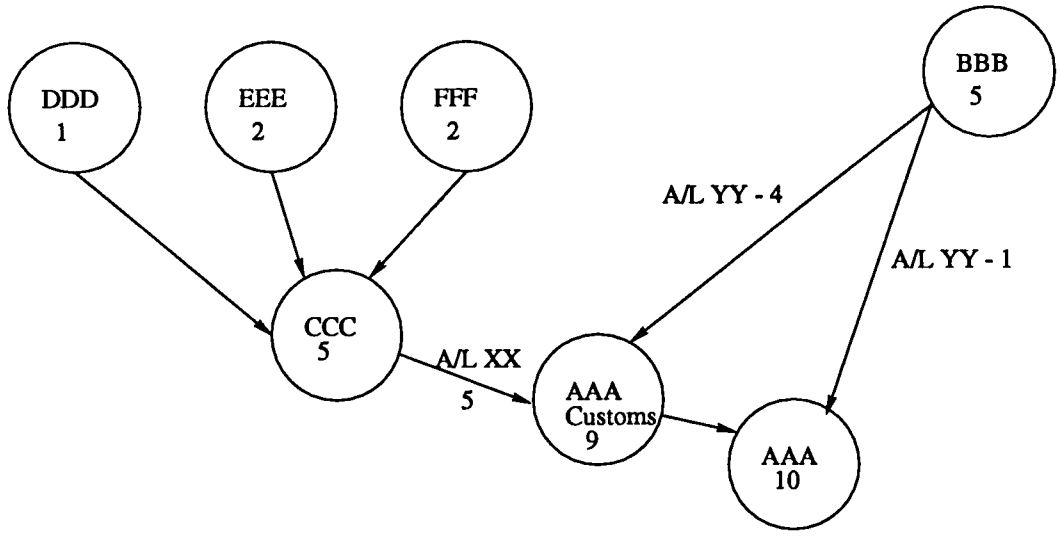
As you requested, I am attempting again to explain what I call my ''common sense'' method of pattern analysis. Before the design if ''frozen'', we should also explore ''textbook'' methods of pattern analysis to see if improvements can be added to my method. In any case, a pattern analysis process must be ''built in'' to your program design.

I identify patterns by first focusing on a group of reports, then searching for elements which are common to most of those reports. I may, for example, look at ten reports of pilfered mailbags delivered to the air mail facility at airport AAA in February 1996. (Those ten reports would be selected based on the fact these pilferages are recurring with high frequency.) Since Airline XX transported the mail in 5 reports and Airline YY transported the mail on the other 5, there is no pattern recognized for an airline. No pattern is recognized for other airports, since 5 of 10 reports the mail came from Airport BBB, and in 5 of 10 reports it transited Airport CCC. Another possible point of loss reported on 9 of 10 reports was ''AAAC'' (the code for the customs warehouse of Airport AAA. -- In this example, all 9 mailbags had arrived at AAA Airport on Saturday afternoon, so they had to be stored in the customs warehouse until they could be delivered to the air facility on Monday morning). My conclusion is that 9 of 10 pilferages probably

occurred at the AAA cutoms warehouse, and that the tenth pilferage is unrelated. I would, therefore, recommend attention be given on weekends at the customs warehouse.

In regard to purging of out-dated data, I suggest that in the beginning of every calendar year we purge reports which are older than the prior calendar year. I will oversee the input of reports, so questionable data hopefully will not get into the system.

The example above could be represented using the network graph shown below:



Appendix C

Data Input Format

Table C.1: Column two indicates whether the data is in alpha (A) or numeric (N) format.

Positions	Format	Description
1-6	A6	Origin Exchange Office of the Dispatch
7-12	A6	Destination Exchange Office of the Dispatch
13	A1	Mail Category Code
14-15	A2	Mail Sub-Class Code
16	N1	Last Digit of Year
17-20	N4	Serial Number of the Dispatch
21-23	N3	Receptacle Serial Number within Dispatch
24	N1	Highest Number Receptacle Indicator
25	N1	Registered/Insured Indicator
26-29	N4	Receptacle Weight in Multiples of 0.1kg
30-37	N8	Date Dispatch was Prepared
38-45	N8	Date Consignment made to Carrier
46-53	N8	Date Pilferage was Detected
54-55	N2	Receptacle Event Code
56-61	AN6	Carrier Number 1/Flight Number
62-73	A12	Airstops
74-79	AN6	Carrier Number 2/Flight Number
80-91	A12	Airstops
92-97	AN6	Carrier Number 3/Flight Number
98-109	A12	Airstops

Vita

Linda Dupont was born on January 14, 1968 in Goffstown, New Hampshire. She got her BS in Computer Scince at the University of New Hampshire. After graduating from the University of New Hampshire, she worked as a computer consultant before beginning gratuate study at the University of Tennesse.