1994

AN INVESTIGATION INTO AN EXPERT SYSTEM FOR TELECOMMUNICATION NETWORK DESIGN

PAUL LAURENCE REYNOLDS

PHD

AN INVESTIGATION INTO AN EXPERT SYSTEM FOR TELECOMMUNICATION NETWORK DESIGN

REYNOLDS, PAUL LAURENCE

Award date:
1994

Awarding institution:
University of Plymouth

AN INVESTIGATION INTO AN EXPERT SYSTEM FOR TELECOMMUNICATION NETWORK DESIGN

PAUL LAURENCE REYNOLDS

BSc, CEng, FIEE

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

Doctor of Philosophy

SCHOOL OF ELECTRONIC, COMMUNICATION and
ELECTRICAL ENGINEERING

In collaboration with
BRITISH TELECOMMUNICATIONS PLC

June 1994

For

Helen Cecily Fletcher Sandbach

An Investigation into an Expert System for Telecommunication Network Design

PAUL LAURENCE REYNOLDS

ABSTRACT

Many telephone companies, especially in Eastern-Europe and the 'third world', are developing new telephone networks. In such situations the network design engineer needs computer based tools that not only supplement his own knowledge but also help him to cope with situations where not all the information necessary for the design is available. Often traditional network design tools are somewhat removed from the practical world for which they were developed. They often ignore the significant uncertain and statistical nature of the input data. They use data taken from a fixed point in time to solve a time variable problem, and the cost formulae tend to be on an average per line or port rather than the specific case. Indeed, data is often not available or just plainly unreliable. The engineer has to rely on rules of thumb honed over many years of experience in designing networks and be able to cope with missing data.

The complexity of telecommunication networks and the rarity of specialists in this area often makes the network design process very difficult for a company. It is therefore an important area for the application of expert systems. Designs resulting from the use of expert systems will have a measure of uncertainty in their solution and adequate account must be made of the risk involved in implementing its design recommendations.

The thesis reviews the status of expert systems as used for telecommunication network design. It further shows that such an expert system needs to reduce a large network problem into its component parts, use different modules to solve them and then combine these results to create a total solution. It shows how the various sub-division problems are integrated to solve the general network design problem. This thesis further presents details of such an expert system and the databases necessary for network design: three new algorithms are invented for traffic analysis, node locations and network design and these produce results that have close correlation with designs taken from BT Consultancy archives.

It was initially supposed that an efficient combination of existing techniques for dealing-with-uncertainty within expert systems would suffice for the basis of the new system. It soon became apparent, however, that to allow for the differing attributes of facts, rules and data and the varying degrees of importance or rank within each area, a new and radically different method would be needed.

Having investigated the existing uncertainty problem it is believed that a new more rational method has been found. The work has involved the invention of the 'Uncertainty Window' technique and its testing on various aspects of network design, including demand forecast, network dimensioning, node and link system sizing, etc. using a selection of networks that have been designed by BT Consultancy staff. From the results of the analysis, modifications to the technique have been incorporated with the aim of optimising the heuristics and procedures, so that the structure gives an accurate solution as early as possible.

The essence of the process is one of associating the uncertainty windows with their relevant rules, data and facts, which results in providing the network designer with an insight into the uncertainties that have helped produce the overall system design: it indicates which sources of uncertainty and which assumptions are were critical for further investigation to improve upon the confidence of the overall design. The windowing technique works by virtue of its ability to retain the composition of the uncertainty and its associated values, assumption, etc. and allows for better solutions to be attained.

CONTENTS

LIST OF FIGURES

GLOSSARY OF TERMS

Access Network

The network that connects the customer to the trunk network.

Access Line

A single line between a customer and his local node.

Alternative Routeing

A procedure whereby several routes are searched to complete a connection.

Asynchronous

Non synchronous mode of operation in which the completion of one portion of a call on one channel initiates another.

ATM

Asynchronous Transfer Mode is a method of customer access and trunk network system that allows for capacity in terms of bits/sec to be allocated on demand.

Bellcore

Bell Communications Research, established to represent the seven US Regional Bell Operating Companies. Bellcore was previously the research arm of AT&T.

BHCA

A nodes central processing unit ability to process calls is measured by the number of Busy Hour Call Attempts (BHCA) it is able to handle.

Broadband

A link or channel which has a line rate of 2.048 Mbit/sec or greater.

Busy Hour

The busy hour is the uninterrupted period of 60 minutes for which traffic is at a maximum. In planning the capacity required in a telecommunication network, it is necessary to assess the traffic volume expected during the busiest period. In a Public Switched Telephone Network (PSTN), the traffic tends to reach a peak during the morning of a normal working weekday.

Call Logging

Call logging (Monitoring) is the collection of data, by a process of formal observation of the performance of an existing telecommunication system. It is the information that results from the analysis of this data that is intended to assist the network designer plan improvements in the performance of the system or forms the input to network modelling tools to design a new, replacement network.

Calling Rate

The traffic per customer in the busy hour.

Channel

A channel is a part of a TDM or FDM link that is allocated for one customer's traffic.

Community of Interest

Community of Interest defines the traffic profile between the nodes of a network based upon the telecommunication relationships of the customers using the network.

Concentration

Concentration occurs when the number of inlet channels is greater than the number of outlet channels. Although concentration makes the best use of equipment and circuits by increasing the traffic loading of each channel, there is a consequential chance that calls will be lost.

Cost of Failure

The cost of providing additional network elements to ensure reliability and right quality of service more than that provided to meet the forecasted demand.

Cross Connectors

Cross Connectors provide interconnection using either a set of manually arranged patching frames or a manually operated node, i.e. one that operates under the control of the network operator and which generally maintains its configuration for periods of several hours, days, months or years.

Customer

The source and sink of traffic on a telecommunication network. (see User, these terms are not interchangeable).

Erlang

One channel occupied for one hour is defined as one Erlang.

Extensions

An extension is a telephone access line provided between a customer and a private branch exchange. This is equivalent to a direct access line between a customer and his local node

FDM

The link which conveys its channels over a common path in which the available frequency spectrum is shared is known as frequency-division multiplex link.

Gateway Node

A gateway is a node which acts as the interface between two networks, it is connected by links to transit nodes within its own network and gateways in other networks.

Grade of Service

Grade of service is a practical interpretation of congestion, i.e. the effects of the failure to establish a call due to unavailability of network paths. It is measured in percentage of calls lost for a system.

Green Field

A site or geographic area under investigation that currently has no telecommunication services.

Heuristic

Any method or technique obtained by expertise or 'rule of thumb' activities and used to improve the efficiency, or feasibility, of a problem-solving system. A heuristic algorithm may work well with a variety of problems and may fail on others.

Hierarchy

A telecommunication network arrangement including the linking of transit and local nodes by links to provide an end-to-end switched service. It is usual to establish a number of central nodes that have high capacity links between them. These links are usually multiplexed so that a particular circuit can handle several hundred calls at any one time. The customers are connected to the network by access lines. Each customer has a separate link to its own local node. The use of local and transit nodes forms a two level hierarchy.

Links

Network links interconnect the nodes of a network.

Local and Transit Nodes

Local nodes are thoes to which customers are connected via access lines. Transit notes are those which interconnect all the local nodes.

Loop Networks

In such a network there is no central controlling node but instead an agreed upon protocol by which each node in the network communicates with others. One node may have certain overall functions (such as a sink in a mutually synchronised scheme), but in principle, each node directly communicates with adjacent nodes only.

Marginal Cost

The cost of providing an addition network element less the revenue accrued from its usage.

Matrix Erosion

A process whereby the largest number in a n*n matrix array is progressively reduced in value. Where the largest occupies more than one cell in the array, then each is progressively reduced.

Mesh Networks

A mesh network is one in which there are many links between nodes and many choices of possible routeings for a call. Taken to the limit, every node would be connected to every other, but this tends to be expensive in terms of links.

Multidrop Networks

This is a simple topology in which nodes are positioned along a single link. Such patterns are common in data networks and telephony networks where the cost of the link is a predominate expense in the construction of the network and it is desired to utilise the link rather than invest in additional links.

Multiplexing

Multiplexers enable a number of low-speed circuits to be carried over a single high-speed link. The number of tributaries is equal to the number of channels carried over the multiplexed trunk link. Two types of multiplexing are used in this thesis for network design: Time Division Multiplexing (TDM), and Frequency Division Multiplexing (FDM).

Narrowband

A link or channel which has a line rate of less than 2.048 Mbit/sec.

Network Element

An individual building block that can form part of a telecommunication network. See Figure 0.

Network Topology

The physical arrangement of the nodes and the interconnecting links in the network are dependant upon the requirements of the application and geographical distribution of the customers. Five distinctive network patterns can be found in practical network topologies, these are the star, loop (or ring), multidrop, tree and mesh. In practice, most telecommunication networks are formed of some combination of the topologies.

Nodes

A network node can provide switching, cross-connection, multiplexing or concentration functions.

PBX

Private Branch Exchange is a node located in a business premise and is use solely for the interconnection of private network traffic.

Plesiochronous

A method of maintaining (near) synchronisation by the use of highly accurate clocks at each node in the network.

Point Source Traffic Matrix

A matrix in which each cell in the matrix contains the total source traffic for a particular geographic location. Communities of Interest are not identified

Port

A port is point of a node that is allocated for one customers traffic.

Private Circuits

A link or channel sold by a public telecommunication network operator to a customer.

PSTN

Public Switched Telephone Network is used to describe the public telephone system including the customers equipment, access network, local nodes, transit nodes, links and hierarchy that makes up the network.

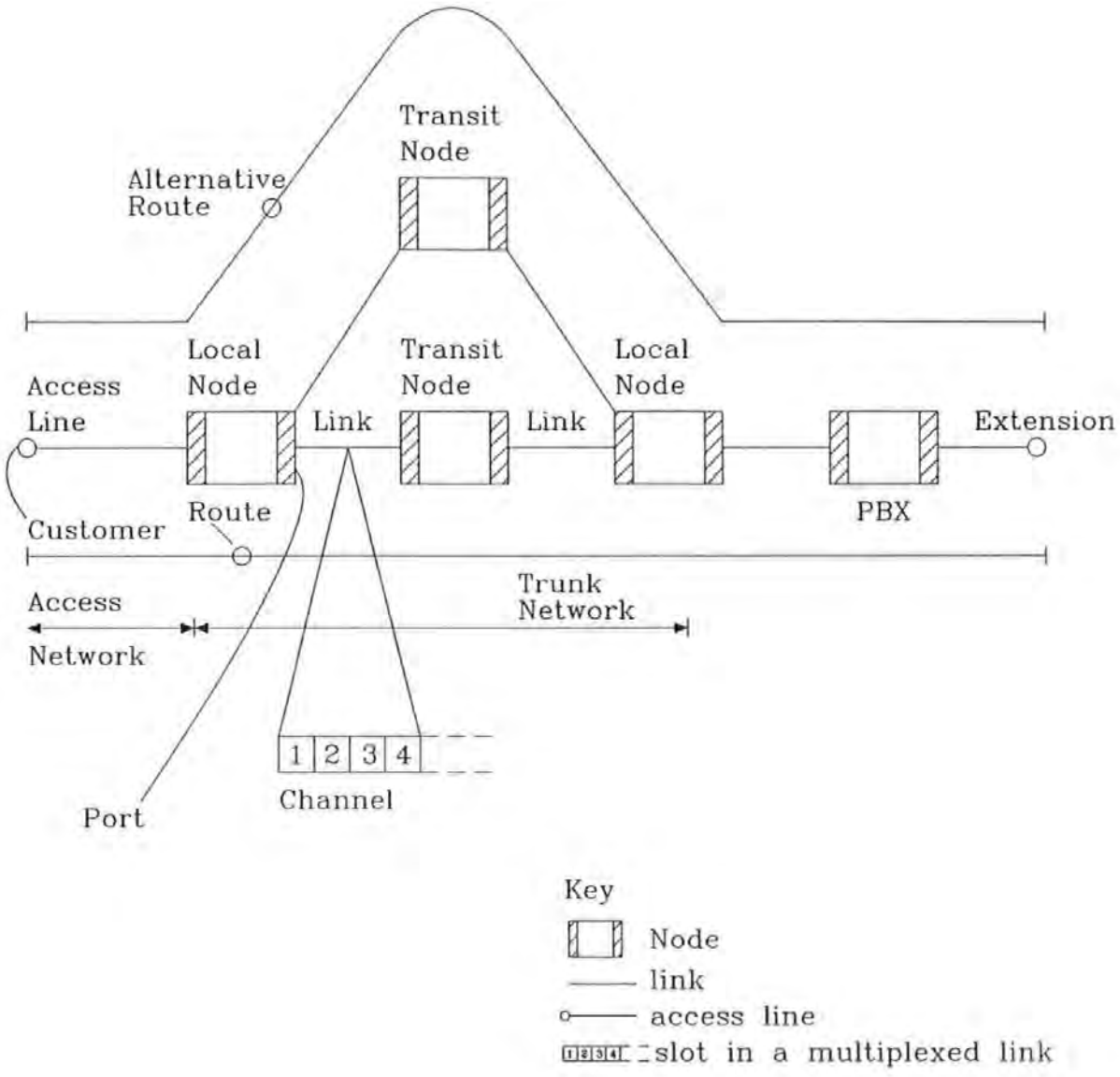


Figure 0 Relationship Between Elements of a Network

Route

A route is a series of links that form the path between two customers.

Routeing & Alternative Routeing

There are many possible routes for a particular call through a telecommunication network, and the combination of the transit nodes, local nodes and links to make the call is known as its routeing. The routeing that uses the least number of links is known as the minimum span routeing for a call; other routeings are known as alternative routeings.

SDH

Synchronous Digital Hierarchy is a transit link network that is maintained in synchronisation. It will replace the plesiochronous transit link networks currently employed by all the worlds Telco's.

Star Networks

This type of network is characterised by having a central node at which a controlling Central Processing Unit handles all the transit functions (including all the switching functions for Remote Peripheral Units) involved in call routeing. All non-local telecommunication between customers takes place to and from, or by way of, the centre.

Switches

Switches provided interconnection between an equal number of inlet and outlet channels. Switching may be achieved with or without traffic loss and form the most significant part of the telecommunication network. The total traffic capacity of a node's Central Processor Unit is measured in Busy Hour Call Attempts. Switches can either be local or transit depending upon functionality.

Synchronous

An arrangement for operating a network at a common (synchronised) clock rate.

TDM

Time-division multiplexing is the process whereby the available occupation time is shared between channels on a common link.

Telco

Telephone Company or Network Operator providing services and network capacity to customers.

Teletraffic

Telephony and data traffic.

Traffic Matrix

Each cell in a matrix contains the source to destination traffic value. From the matrix total terminating, total originating, total inter node traffic and communities of interest can be identified.

Traffic Theory

The theory of call analysis used in the planning of a telecommunication networking which the network designer optimises the equipment and facilities needed to support the requirements of customers. Traffic theory requires the consideration of the average behaviour of customers of a system and relies upon the fact that, in a large enough population, the random behaviour of individuals tends to present a consistent pattern. Traffic theory includes the consideration of the following: number of call in period (n); the holding time of each call (h); the duration of the calls (T) and the rate at which the calls arrive (a).

Traffic Volume & Average Traffic

Traffic volume is defined as the amount of traffic over a given time. If in T seconds n calls are made and if the duration of the calls are h_1, h_2 ... h_n seconds, then the utilisation of the system is:

$\sum_{i=1}^{i=n} h_i$ and the average traffic in a given time is expressed as $E = \frac{\sum_{i=1}^{i=n} h_i}{T} Erlangs$

Traffic

Traffic defines the utilisation or capacity of a telecommunication system and is described in terms of the number of calls in a specified period. Traffic is measured in Erlangs.

Transit Node

A node in a telecommunication network that acts for a particular call as an intermediate node, i.e. not the first or not the last node in a routeing.

Tree Networks

A tree network involves a pyramid hierarchy of nodes in which the highest node (gateway node) in the network appears to be the apex of a pyramid. The lower level nodes of the pyramid act as traffic collectors (local nodes) or distribution (transit nodes) for all communication to and from the other levels.

Trunking

The method of roueting a call through a network node.

Trunk Network

The network of links inter-connecting the nodes in a community group is called the trunk network.

Trunks

A combination of link channels and node ports involved in a call.

User

Within this thesis, a User is defined as the operator of the expert system. (see Customer).

ABBREVIATIONS

AI	Artificial Intelligence
ATM	Asynchronous Transfer Mode
BH	Busy Hour
BHCA	Busy Hour Call Attempts
bit/sec	bits per second
BMA	Bangkok Metropolitan Authority
BP	British Petroleum PLC
BT	British Telecommunications PLC
CCITT	Consultative Committee for Telegraph and Telephony
CCS	Cross Connect Site
ccs/h	call seconds per hour
CF	Confidence Factor
CPI	Consumer Price Index
CPU	Central Processor Unit
ES	Equipment Synchronisation
FDM	Frequency-Division Multiplex
GAS	Autonomous Study Group
GDP	Gross Domestic Product
GOS	Grade of Service
GRP	Gross Regional Product
HDB	High Density Binary
HQ	Head Quarters
IDA	Integrated Digital Access
IMF	International Monetary Fund
ISDN	Integrated Services Digital Network
ITU	International Telecommunication Union
JICA	Japanese International Co-operative Agency
kms.	kilometres
MA	Municipal Areas
NESDP	National Economic and Scientific Development Plan
NSO	National Statistics Office
NSP	Network Synchronisation Plan(ing)
NTU	Network Terminating Unit
PBX	Private Branch Exchange
PC	Primary Centre (US terminology for a transit)
PCM	Pulse Code Modulation
PDH	Plesiochronous Digital Hierarchy

PF	Penetration Factors
PM	Primary Master
ppm	parts per million
PSTN	Public Switched Telephone Network
RSU	Remote Switching Unit
SDH	Synchronous Digital Hierarchy
SDR	Special Drawing Right
SDV	Sanitary Districts and Villages
SM	Secondary Master
SR/B	Speculative Residential/Business
SU	Synchronisation Unit
TDM	Time-Division Multiplexing
Telco	Telephone Company
TM	Tertiary Master
TOT	Telephone Organisation of Thailand
TSO	Time Slot Zero
TU	Timing Unit
US	United States of America
2B+D	Two Primary Channels (B) and a Secondary (D) Signalling Channel

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to:

• **Peter Sanders**, my Director of Studies, without whose enthusiasm, understanding and support the research would have never been completed.

• **Colin Stockel**, my Supervisor, whose guidance and sensible advice brought both direction and realism to my research.

• **David White**, my BT manager back in 1987, who had the foresight to see that I would able to complete the programme of work whilst still in full time employment and supported my application for sponsorship.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was financed with aid from and carried out in collaboration with, BRITISH TELECOMMUNICATIONS PLC.

Relevant scientific seminars and conferences were regularly attended at which work was presented: external institutions were visited for consultation purposes and papers prepared for publication.

Signed .....

Date ..... 1 June 1994

CHAPTER 1

TELECOMMUNICATION NETWORKS

1.1 Introduction

The world's telecommunication network is the largest human construction of all time. It is also one of the most complex, handling in excess of 1000 billion calls every year plus an immense amount of data. Today there are more than 800 million telephones in use in some 230 countries around the world, most of which can be dialled direct, virtually without censorship or control. Its very existence is an important element of our personal freedom and there is little doubt that telecommunications is a fundamental requirement for sustaining modern civilisation. [1]

Although the concept of telephonic communication existed at least 300 years ago, the practical realisation is not yet 120 years old. From that time the progressive penetration of telecommunication services into our lives has grown exponentially and customers of telecommunication network providers are becoming more and more demanding, particularly for services that represent good value for money. The quality with which such services can be offered, and their cost of provision, crucially depends on the structure of the telecommunication network.

Customers are usually indifferent to the structure of the network that provides a service, except in so far as it affects costs and quality. The network operator cannot afford the luxury of such indifference, but must understand exactly how the structure of the network impacts on those key parameters of the product that are presented to a customer.

The term network is invariably used in the singular, however, it may consist of many smaller sub-networks such as telephone, data, visual, intelligent, open service, etc. each having a structure comprising network nodes, network links and customers. A common analogy of a telecommunication network is that of a house which is situated on a private road leading to a public street that connects to the trunk roads. Then several trunk roads in turn lead to motor-ways that are fully interconnected. A road network is very much the same as a telecommunication network: there is a technological hierarchy and a traffic rule base that requires the utilisation of similar traffic analysis techniques.

1.2 Telecommunication Network Architecture

A telecommunication network can be viewed as separate layers of functionality. The layering helps to separate concerns and manage the complexity in the network by breaking down functionality through defined interfaces which support the necessary interconnection and interoperability requirements of a network environment that is diverse and varied in nature.

Figure 1 shows an architectural model of a telecommunication network.

At the model's lowest level are the customers; their interface to the transport layer and higher levels for the presentation and connection to services is by way of the local access network.

The two transport layers, incorporating network node and link elements, are responsible for the efficient connection and transport of services. The Nodal Layer is

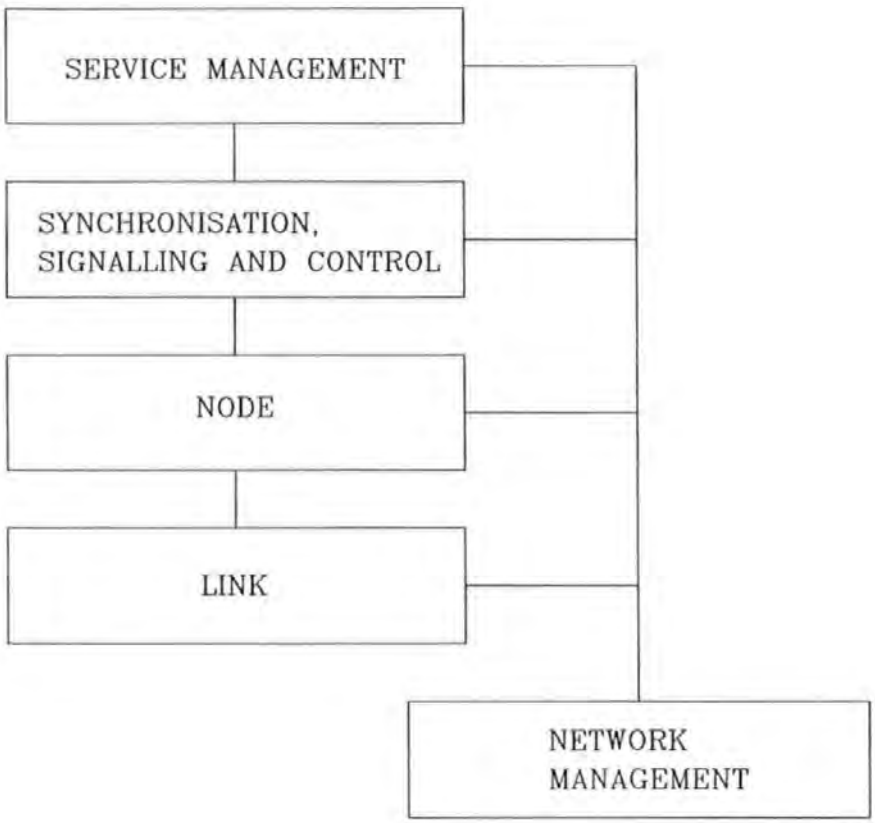


Figure 1 Architectural model of a telecommunication network

in the form of elements based upon 64 kbit/sec 'time-space-time' switching matrices, supporting 64 kbit/sec ports, when connected to the local access network, and 2 Mbit/sec ports when connected to the link network; its elements are responsible for the switching, concentration and multiplexing of telephony, data packet, mobile communications, private circuit, etc. Some transport elements, such as Synchronous Digital Hierarchy (SDH) equipment, consists of both node and link functions and are referred to as Trans-Nodal elements. The Link Layer is traditionally in the form of links comprising multiples of 64 kbit/sec channels. These fixed bandwidth channels are based upon Plesiochronous Digital Hierarchy (PDH); it is expected that this technology will eventually be replaced by variable bandwidth systems called Asynchronous Transfer Mode (ATM).

At the top of the model is the Service Management Layer which incorporates all the network infrastructure necessary to support the creation, provision, monitoring, maintenance and billing of customer's services. Network management forms the network 'glue' that binds the various levels together.

Telecommunication network design is traditionally associated with the two transport layers of the network architecture.

1.2.1 Network Nodes

Network nodes can provide switching, cross-connection, multiplexing and concentration functions. These functions are provided in combinations appropriate to the purpose of the node. Nodes either support the real-time, on-demand

establishment of calls by customers, or the make for the efficient use of resources to provide private network services.

Switching takes place where connectivity is provided between an equal number of inlet and outlet ports. Switching may be achieved with or without traffic loss and they form the most significant part of the telecommunication network. Switches have evolved from manual-boards with operators manipulating plugs into sockets, through step-by-step electro-mechanical and electronically controlled crossbar systems, to today's time division digital switches. In contrast to the early electro-mechanical systems that were constrained to fixed maximums of customers and ports, the digital switch can be configured with any ratio of customer lines to network ports as long as the total traffic capacity of its Central Processor Unit (CPU) is not exceeded.

Cross-connection facilitates interconnection using either a set of manually arranged patching frames or a manually operated switch, i.e. one which operates under the control of the network operator and which generally maintains its configuration for periods of several hours, days, months or years.

Multiplexing enables a number of narrowband channels (64 kbit/sec) to be carried over a single wideband link (2.048 Mbit/sec and above). Its chief use is to reduce link costs by greatly improving their usage and to increase the manageability of a network by reducing the number of links to be monitored and maintained.

Concentration occurs when the number of inlet channels is greater than the number of outlet channels. Typically, a customer uses a telephone line for only a small percentage of the time even in the busiest period of the day, (for example, on the BT

network, for about six per cent of the busiest hour of the day (BH), allowing for both originating and terminating calls). Concentration can boost link usage to around 60 to 70 per cent. Although concentration makes the best use of equipment and links by increasing the traffic loading, there is a consequential chance that calls will be lost.

Nodal elements can take many forms but are grouped into two main technology groupings; narrowband and broadband. Narrowband nodal elements are based upon n*64 kbit/sec channel manipulation. Telephony is all at 1*64 kbit/sec, videoconferencing requiring 6*64 kbit/sec (384 kbit/sec), Integrated Services Digital Network (ISDN) - '2B+D' at 3*64 kbit/sec and Packet Switched Frame Relay at 1*64 kbit/sec. The nodal elements use switching matrices operating at the standard 64 kbit/sec; customers terminal equipment compensates for the differing delays experienced by individual channels forming the n*64 kbit/sec service. Broadband nodal elements are based upon n*2 Mbit/sec channel manipulation. SDH cross-connection and ATM multiplexing are the two main broadband nodal elements operating at 75*2 Mbit/sec (155 Mbit/sec).

1.2.2 Network Links

The functionality of links is simple, by comparison. Links provide a group of channels between specified nodes. In general the link can convey its channels over physically separated paths, i.e. cables, over a common path in which the available frequency spectrum is shared, known as frequency-division multiplex (FDM); or over a common path in which the available occupation time is shared, known as time-division multiplexing (TDM).

Currently, all new networks designed use only TDM digital transmission links between nodes. This TDM is based upon PDH, which multiplexes 64 kbit/sec channels to form link capacities of up to 565 Mbit/sec, i.e. 7680*64 kbit/sec. Link technologies include optical fibre, radio, coaxial and copper wire. PDH has a complex frame structure which, at each stage of multiplexing, has overhead bits added for frame alignment and identification. At level of 140 Mbit/sec and above it is impossible to identify any of the 2 Mbit/sec tributaries without demultiplexing down to that level. SDH has link gross bit rates of 155 Mbit/sec, 622 Mbit/sec and 2.5 Gbit/sec. SDH offers the ability to identify any 2 Mbit/sec tributary in a 155 Mbit/sec frame and any 155 Mbit/sec tributary in a 2.5 Gbit/sec frame without resorting to demultiplexing.

1.2.3 The Customer

In an increasing competitive environment, customers will view price as a major factor in marketing choices between network operators. This is achieved by the provision of cost optimised networks providing the quality of service perceived by the customer as normal. Different market sectors have different expectations of network quality. For large business customers, reliability and resilience are prime requirements. For smaller customers, fast provision of services is the prime requirement. The customer is both the source and sink of traffic to be carried by a network: his traffic profile is dependent upon many socio-economic attributes; for example, their cultural background and status within life, the function of the organisation in which the customer is employed, etc.

The usual method of quantifying the traffic generated by customers is either by a process of call logging, i.e. call monitoring, on existing telephone systems or by market survey techniques, i.e. interviewing new customers. In addition, the network designer has to resort normally to a series of heuristic rules to develop the traffic profile, which in turn forms the traffic matrix, the basic starting point for all current computerised network design tools.

It is the customer and his requirements that will drive the evolution of telecommunication networks, their size, quality, reliability and flexibility and not the technology of the nodes and links. Furthermore, it is the customer who will provide the network operators with the greatest uncertainty reflected in their lack of understanding of his needs. A wide variety of services have been made available in recent times but a relatively few have been commercially successful. This new customer driven environment will require networks that are just adequate to meet their needs. The problem is that newer technologies are only cost efficient if manufactured in large quantities and used in networks that are highly 'tuned' to user demand.

1.3 The Network

There are two categories of telecommunication networks, Circuit Switched and Packet Switched and two classifications of availability, Public and Private.

1.3.1 Public Networks

Public telecommunication networks provide communication between all customers who subscribe to membership of that network. A public network may cover an entire country, or just a specified region, or class of customer within a country. There may be several public networks serving one country. Where this is the case, the networks are interconnected to enable the appropriate degree of access between customers on the various networks. Public networks may be owned and operated by either private companies or government departments.

The ubiquitous telephone networks throughout the world are the prime examples of public telecommunication networks. These are commonly known as Public Switched Telephone Networks (PSTN).

In the UK, besides British Telecommunications and Mercury, networks are now being provided by numerous cable franchises and some ex-public utilities.

1.3.2 Private Networks

Private telecommunication networks provide communication between members of a single organisation. In this sense, the composition of an organisation may be just one private company, a conglomerate of several companies, or government departments. The limits of such organisational groupings that may operate a private network are defined by the regulatory regime in force.

Although the domain of a private network is thus restricted, such networks can extend across a whole country or even across many countries - global private networks. Indeed, there are many possible ownership arrangements for private networks. Links that interconnect the nodes on the private networks are usually provided and operated by a public network operator in the form of 'private circuits', although in certain circumstances, links may be provided by the private network operator itself or a third party.

However, the nodes of the private network are usually owned, provided and operated by the private organisation. In some instances, the private organisation may elect to use a public network operator or a third party to provide and operate the private network on its behalf. In this case the nodes and links of private networks are dedicated to the customer's organisation.

Since many employees of the organisation are associated with each other in an office environment, factory, hospital, etc., with most of the communication needs being between fellow employees rather than the outside world, it is often uneconomical to connect each telephone to a node that may be several miles away. Nodal equipment is therefore placed in the business customers premise and may be configured so that it is located in one building and connected, by direct cable or other transmission systems, to other remote locations of the business customer. This type of nodal equipment is called a private branch exchange (PBX). In some cases, customers may link their PBXs in different locations by 'private circuits' to provide a private telephone network for their own exclusive use. This is economic if the amount of traffic between the different locations justifies the cost of leasing links from the

network operators instead of paying for calls over the PSTNs. These links are now normally digital having capacities of multiples of 64 kbit/sec and 2.048 Mbit/sec.

1.3.3 Circuit Switched and Packet Switched

Two uniquely different types of telecommunication networks are possible, these are packet switched and circuit switched.

In a telephony network, a customer is able to talk to another by dialling a connection via the local node, and then by links (possibly including other nodes) to the required customer. Once a route has been established between the two customers, it remains available until the call is finished. This is an example of circuit switching.

There are, however, switching systems in which links are used but not allocated exclusively for the duration of a particular call. Such systems are more often used in the transport of data. In a packet switched system, data is transmitted in blocks which contain identifying information, allowing the packet to be associated with a particular call to a particular customer: as they pass over the links, the packets forming the elements of the as message may be interspersed with packets related to entirely different calls, destined for other customers. No continuous available route exists between the two customers.

The distinction between a packet switched system and a circuit switched system is determined largely by their design of the node. The former stores information until it can establish a route whilst the later establishes a physical trunk between its input and output for the duration of the call.

1.4 The Arrangement of the Thesis

Chapter 2 overviews the traditional approach to telecommunication network design. It examines graph theory and its application to the design process. The chapter concludes with an analysis of the problems of using the traditional approach in practice.

Chapter 3 introduces the philosophy of intelligence and follows an argument through to a definition of expert systems. The ability of expert systems to handle uncertain information is discussed.

Chapter 4 discusses how an expert system is developed with emphasis on telecommunication network design.

Chapter 5 starts the new work by detailing three new network design algorithms developed by the author for telecommunication network design.

Chapter 6 continues the new work, with reference to dealing with uncertainty in an Expert System and details the new Uncertainty Windowing technique.

Chapter 7 details the implementation of an Expert System based upon the new work and gives results of a design for the City of Bangkok.

Appendix 1 adds details to the Uncertainty Windows principal whilst Appendices 2 and 3 give case studies supporting the work in the main body of the thesis.

CHAPTER 2

THE TRADITIONAL APPROACH TO NETWORK DESIGN

2.1 Introduction

This chapter reviews the ways in which telecommunication networks are currently designed. The reader is first taken through the traditional design methodology, illustrated by way of a simple example. The methodology is then detailed starting with estimating the number of customers on the network and their respective traffic requirements; the chapter moves on to the principles of graph theory and the application of the theory to the design of telecommunication networks. The chapter concludes with an analysis of the problems of using a traditional approach in practice, in particular with the aid of computer systems, and the associated effects of data uncertainty.

2.2 Fundamental Network Design Issues

Node and link equipment are situated at sites which facilitate easy access to cables which radiate out along ducts under streets and highways to the customers.

In existing networks, the position of these sites has arisen from historical, grounds; economic trade-offs between cable conductor sizes, building fixed costs and the then traffic densities. The cable conductor size determines the impedance characteristics of the local line and this in turn determines how far customer's equipment can be situated from the node.

In the case of 'city centre' nodes, such distance problems are not generally the main issues but it is simply the sheer number of customers that it is desirable to connect to the node.

The smallest networking node is capable of carrying large amounts of traffic. It has a substantial fixed cost irrespective of its traffic carrying capacity, due to power supplies, announcement equipment, etc., so the fewer the nodes that are provided the lower the total cost of the buildings and equipment. Conversely, if only one node were used it would not give an optimum link cost where the average distance between the customers and the node should be as small as possible, which demands a corresponding large number of nodes. In practice, the problem usually reduces to one of providing a local node per community. Furthermore, if a community served by a single node still has many of its access lines very long, it could result in the cost of the access network being excessive; it is then economical to divide the community into separate areas, each served by its own local node. The cost of providing additional nodes is more than offset by the reduced cost of the required shorter access links.

Hausen-Tropper [2], of the American telecommunication research division Bellcore, states that the objective of the network design is *to determine how many communication nodes are necessary to support the traffic, and to determine how best to connect these nodes; among the design parameters are:*

communication link capacity;

cost of the network elements;

connectivity (Community-of-Interest);

security or fault tolerance.

The design criteria will vary according to customer requirements and circumstances but is usually one of the following:

cost;

sensitivity to traffic loading;

survivability.

The network design process is interactive. Design methodologies need to accept that there could be data missing and that solutions have to fit to the practical constraints of the availability of plant, accommodation and unknown customer demands.

Fundamental structure issues to be considered in determining optimum network can be illustrated by examining the interconnection need of a simple five-node network. The network topology options available to the designer range from non-hierarchical, figure 2, through to hierarchical shown in figure 3; a variety of hybrid network options such as star, mesh, and loop are also possible, these are shown in figure 4. In the case of a small network, it is usually sufficient to utilise a fully interconnected or 'mesh' type arrangement; every node being connected to every other. The number of direct links with n nodes is n(n-1): thus as n increases the number of links in a mesh increases rapidly. The effect is of reducing the average traffic level per link and hence a less efficient utilisation of the link. At the other extreme, in the 'star' arrangement, the minimum number of links connecting n nodes is (n-1). There are a large number of topologies that can support this arrangement ranging from a simple-star (two level hierarchy) or multi-star (three or more level hierarchy).

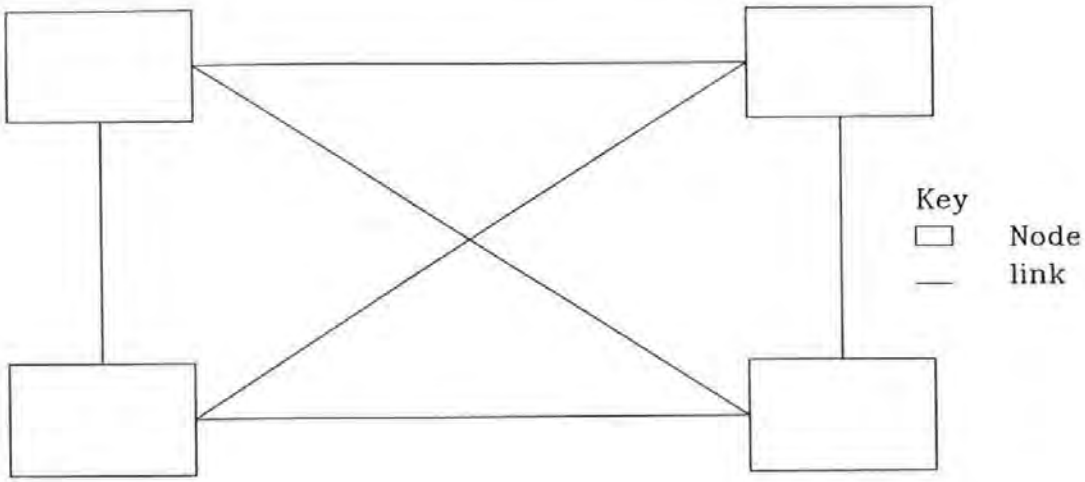


Figure 2 Non-Hierarchical Network

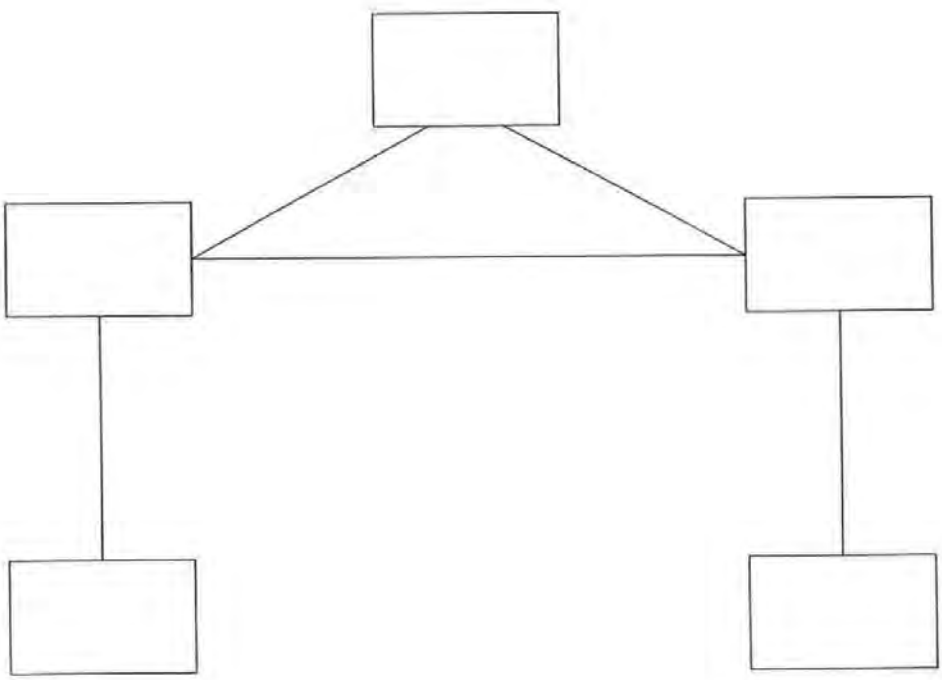


Figure 3 Hierarchical Network

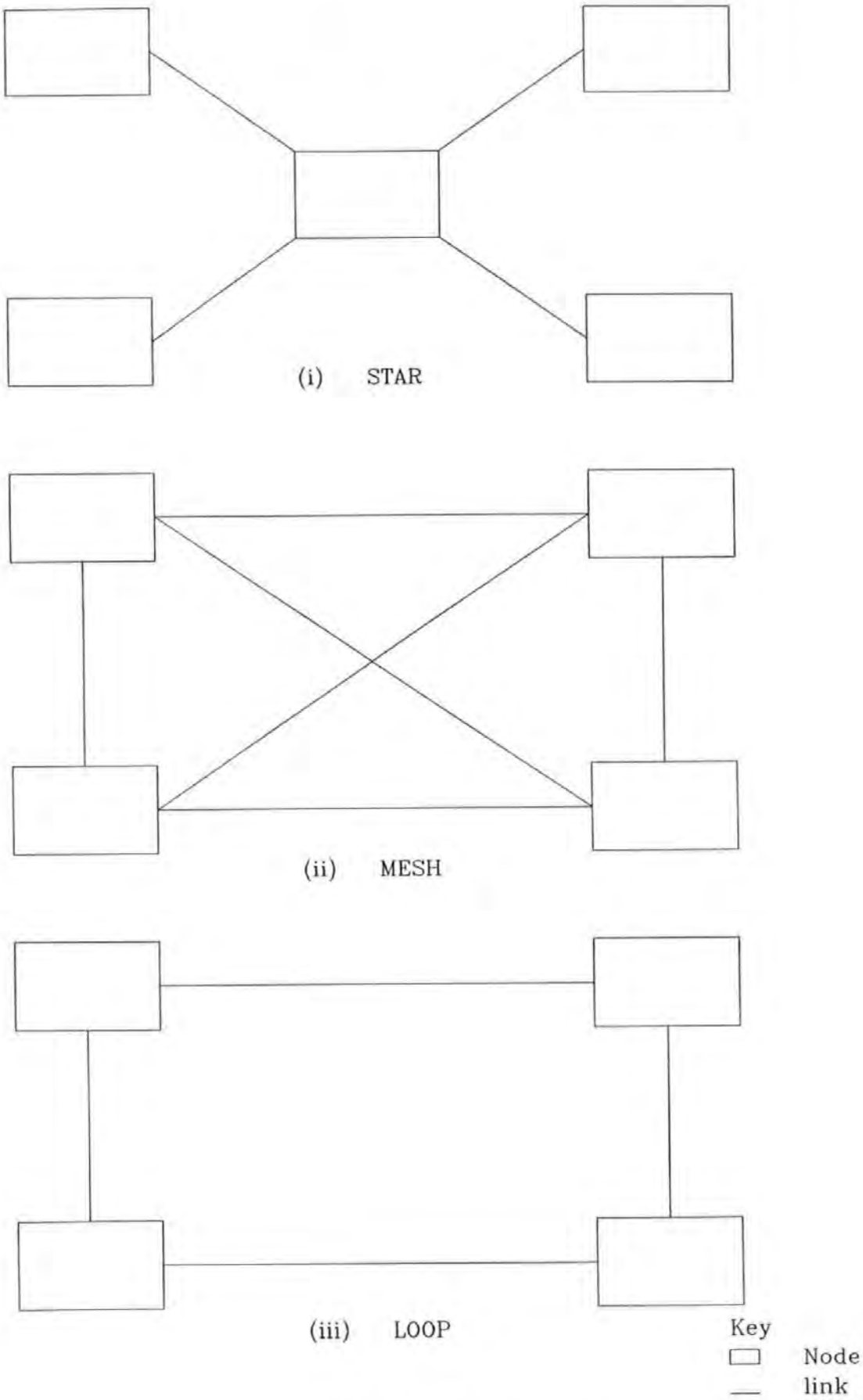


Figure 4 Hybrid Networks

The correct design structure becomes more important as the number of nodes increases, for example, the current BT network would require almost two million links to interconnect fully the 6300 nodes on its public network. If the nodes were connected by such a network, it would lead to a very high-cost interconnecting network. However, an alternative would be to interconnect fully, say, 100 nodes, with all the rest being connected to only one of the 100 fully interconnected ones. This yields a much lower-cost network solution.

The hierarchical pattern facilitates a number of other network requirements including synchronisation and signalling. The network links inter-connecting the nodes in a community group are called the trunk network and the nodes which interconnect all the local nodes, besides its normal local functions, are called transits.

2.3 Traditional Network Design Methodology

The initial stage in the design of a telecommunication network is to determine the base data required to develop the traffic requirements of the network. One such requirement is the determination of number and type of customers that will use the network.

2.3.1 Estimating the Base Data

Market measurement and forecasting techniques are used in developing quantitative estimates of customer demand and their traffic requirements. There are many techniques available to the network designer, but selecting the most suitable method depends on:

accuracy of past data;

availability of data;

final application of forecast and degree of error that can be tolerated;

time available to prepare the forecast.

If the network designer is aware from previous work that the data available contains some inaccuracies which can not be eliminated, then it is wasteful to expend effort on precise market mathematical techniques and subsequent network modelling. Similarly, it is unnecessary to use a sophisticated technique requiring a run of data when only a limited set of data are available.

It is important to make the best use of the data, but at the same time to keep within limits of acceptable accuracy.

Four estimation techniques are available to the network designer:

opinion surveys;

time series analysis;

cross sectional analysis;

call logging (monitoring).

2.3.1.1 Opinion Surveys

Opinion surveys are particularly suited to telecommunication forecasting, especially in areas where the market does not lend itself to any form of mathematical analysis; e.g. estimating telephone demand in a yet to be developed infrastructure. The chief

weakness of the opinion survey process is that it is highly subjective and no matter how many adjustments are made for bias it may be severely criticised on these grounds alone.

2.3.1.2 Time Series Analysis

A 'time series' is a set of data which shows the behaviour of a variable, e.g. telephone penetration, over a regular interval of time. The objective of time series analysis is to identify measurable regularities, patterns and correlations in the past data which can be assumed to recur in the future, and thus provide some clues for forecasting the future fluctuations of the series.

2.3.1.3 Cross Sectional Analysis

A 'cross section' is a set of data which shows a trend, e.g. customer calling rates, at a fixed point in time. The objective of cross sectional analysis is to identify measurable regularities, patterns and correlations in data from either various countries or models which can be assumed to occur in the same time period in a new environment.

2.3.1.4 Call Logging (Monitoring)

If a telecommunication network already exists it is possible to monitor the traffic flowing on that network. This is achieved by the connection of monitoring equipment at each node of the existing network and the recording of the signalling information and call durations. Logging will allow the network designer to know:

where the traffic originates;

where the traffic is destined;

what type of traffic it is.

For a replacement network to be constructed in a cost effective manner, a complete understanding of the traffic flows in the existing network is required; unless the logged traffic is properly analysed, gross inefficiencies that lie hidden within the network can pass unnoticed, what is more, these inefficiencies tend to be promulgated through successive network designs. Traffic statistics produce by logging are also likely to be in an idiosyncratic format, i.e. randomly listed and unprocessed, requiring considerable 'massaging' and interpretation.

2.3.2 Processing the Data

Good forecasting of customer demand, and hence their traffic profile, is fundamental to efficient network design. The need arises because the logistics of telecommunication equipment provision are such that there is a considerable time lag of the design lead-time between identifying a requirement and satisfying it. It is thus vital to forecast demand sufficiently far ahead for action to be taken and requires information on when resources are likely to be exhausted and how much should be ordered for replenishment to cover their design period.

After all the base statistics, logging data and economic and socio-economic data has been collected it is then necessary to convert this data into a form that can be used to determine the optimum size, location and traffic collecting boundaries of the local nodes. From the growth of connections for each node, its originating traffic is

traditionally derived by applying appropriate calling rates to the mix of residential and various business sectors served by the node. It is then necessary to determine traffic distribution between the nodes according to the community of interest between customers served by the nodes. The total network can then be dimensioned and costed.

The resultant design should provide:

satisfactory service to the customer;

low cost;

flexibility to meet traffic variations and deviations from forecast;

high resilience against breakdown and unexpected changes in traffic profiles.

Because these characteristics are incompatible, there is a need for what is known as 'good' design. This involves the study of combinations of node and link plant together with the traffic routeings to choose the solution that best meets the optimisation criterion. The criterion is generally the minimum cost solution to meet certain quality of service criteria for a given traffic demand forecast; another optimisation possibility includes maximising capacity for a given cost.

The determination of the number, location and traffic collection areas of nodes is an example of network optimisation. For the access network, a small number of nodes results in large catchment areas; the cost of the local access network linking the node to customers is therefore high because not only is the average length of a link long, but cable costs are high to meet electrical resistance and signalling limits. However,

the cost of node equipment, sites and buildings and the transit link network is comparatively low because of scale economies.

If the number of local nodes is increased, the cost of the access network decreases because of shorter, less expensive links are needed, at the same time, the cost of local nodes, sites and buildings increases. The optimum solution is a compromise between these two extremes.

The design process usually comprises the application of rules and procedures to sub-optimise various network elements such as the transit link network, traffic routeing plan and local cable network which, when brought together, give a reasonably well dimensioned network. An essential element of the design is in determining the most economic traffic routeing. This is carried out by converting the traffic between all nodes into channels routed direct or by way of intermediate nodal points (transit nodes) taking account of routeing and link constraints. A traffic route matrix of channels between all node points can now be assembled and, from this, the physical link plant layout is determined, the total link and node quantities calculated and all finally costed. Figures 5 to 9 illustrate, by a simple example, the various traditional network design stages.

The first stage, illustrated in figure 5, is the derivation of node locations and their traffic collection area boundaries. The location of the node sites is determined by a detailed study of the costs involved in providing access line plant from a number of possible node sites. The optimum locations depend upon the actual and proposed design of cable routes in each nodes traffic collecting area and are obtained by considering possible locations of each node along these routes. The location of the

Node	1	2	3	4	5
1	–	10	20	25	30
2	10	–	23	30	30
3	20	23	–	10	12
4	23	30	10	–	21
5	30	30	12	21	–
	Distance (kms)				

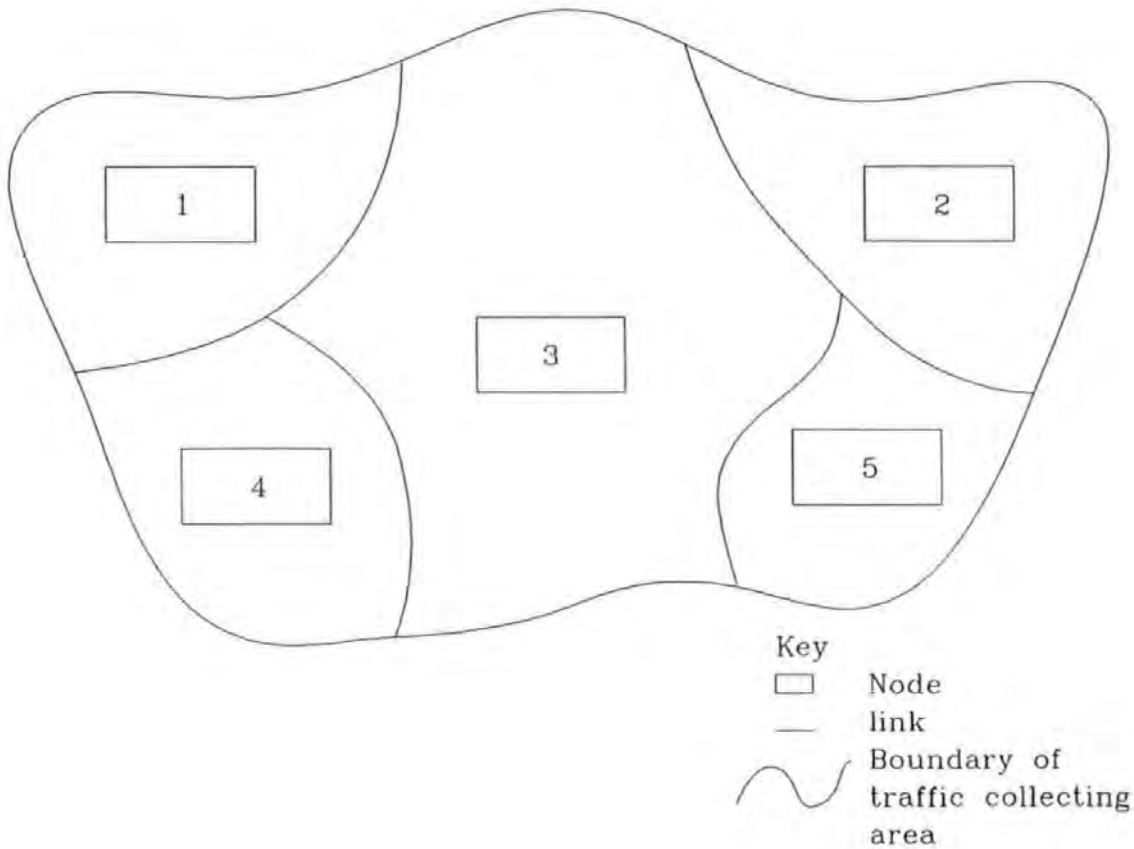


Figure 5 Derivation of Exchange Locations

nodes and the layout of the cable must therefore be done simultaneously. Potential customers are then identified and allocated to their nearest node.

The second stage, figure 6, is one of developing distance and connection forecast matrices for each node area. The process is achieved by technicians walking the streets that comprise the node area and estimating the penetration of telephone in each house in every street. Average calling rates are used to calculate the originating traffic levels.

Figure 7 shows the result of distributing the traffic, based upon communities of interest, into the form of a traffic matrix. In calculating the traffic distribution in a network, typically comprising a central site or head office and remote sites such as factories, stores, etc. it is traditional to use both logging and forecasting techniques to detail traffic distribution and communities of interest.

In the absence of such techniques, it is usual, within BT Consultancy, to use the following set of rules with an aim to constructing the traffic matrix.

Case 1 Central Site (i.e. head office, etc.):

60% traffic is local;

40% traffic is distributed to all other sites of interest.

Case 2 Non-Central Site (i.e. office and non-manufacturing):

70% traffic is local;

Node	Connections	Calling Rate	Originating Traffic (E)
1	1700	0.02	34
2	2300	0.02	46
3	1500	0.03	43
4	900	0.02	17
5	2100	0.03	63
	8500	–	203

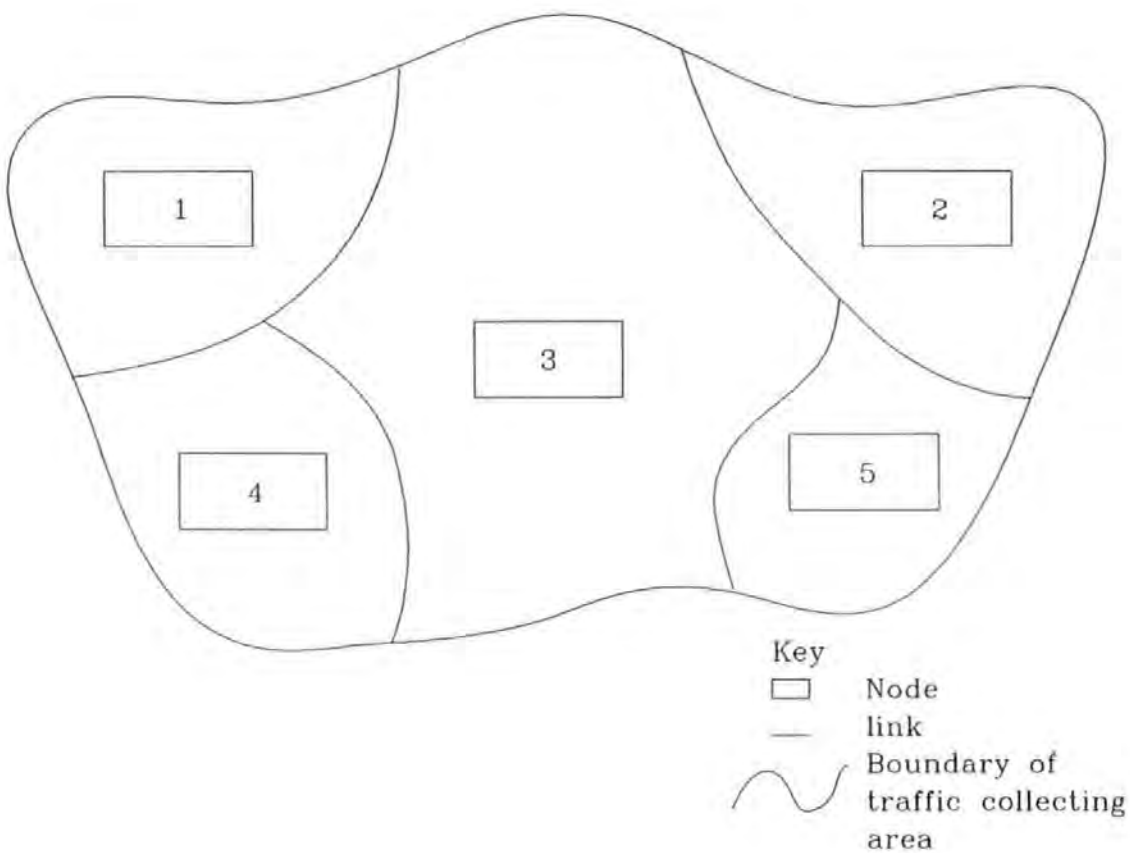


Figure 6 Forecast of Connections and Traffic

Node	1	2	3	4	5	Total Originating Traffic
1	–	20	5	3	6	34
2	18	–	3	4	21	46
3	6	2	–	5	30	43
4	5	3	7	–	2	17
5	4	23	32	4	–	63
Total Terminating Traffic	33	48	47	16	59	203

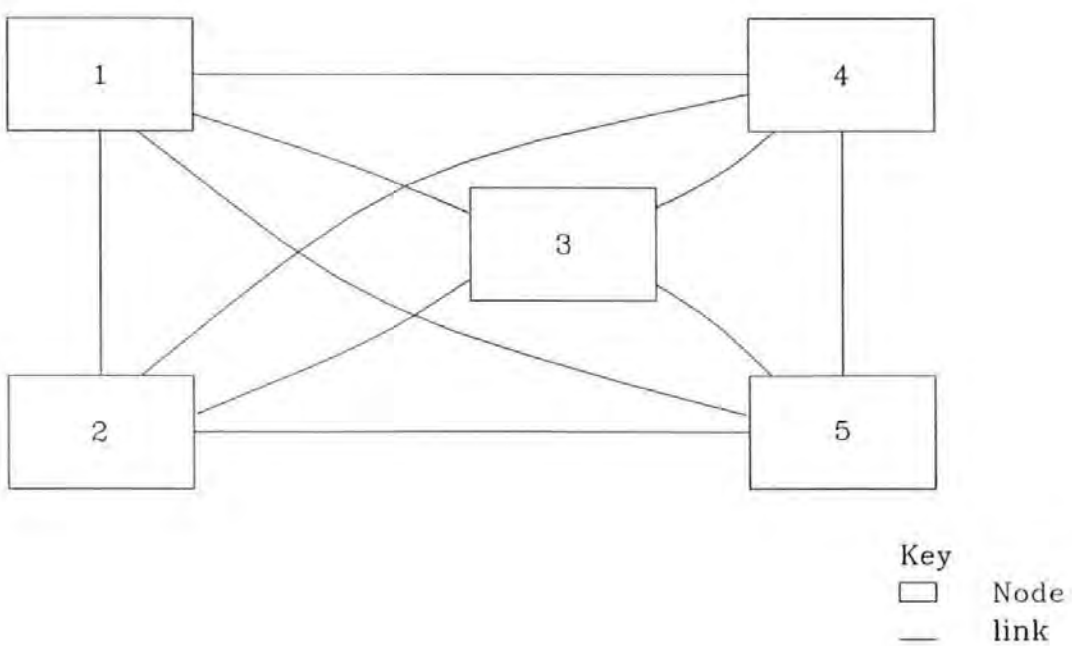


Figure 7 Determining the Traffic Distribution

20% traffic is directed to central function;

10% traffic is distributed to all other sites of interest.

Case 3 Non-Central Site (i.e. manufacturing):

5% traffic is local;

95% traffic is directed to central site.

Case 4 Site with incomplete traffic profile:

reciprocate traffic level.

The third stage of design is the determination of the traffic routeing pattern, shown in figure 8. This involves the comparison of cost of routeing a channel either directly between the two nodes or by way of an intermediate transit node. The economies of scale of the provision of links for combined traffic links can often outweigh the additional cost of transits.

The fourth stage is one of converting the traffic on the routes to channels using graphs relating the two variables for differing grades of service. Figure 9 illustrates the transformation.

The final stage is one of dimensioning and costing the nodes. A node will have fixed costs, e.g. of power supplies, common equipment and announcement facilities, plus a variable cost determined by the number of customer ports, number of links and hence

Node	1	2	3	4	5
1	–	d	t	t	t
2	d	–	t	t	d
3	t	t	–	t	d
4	t	t	t	–	t
5	t	d	d	t	–

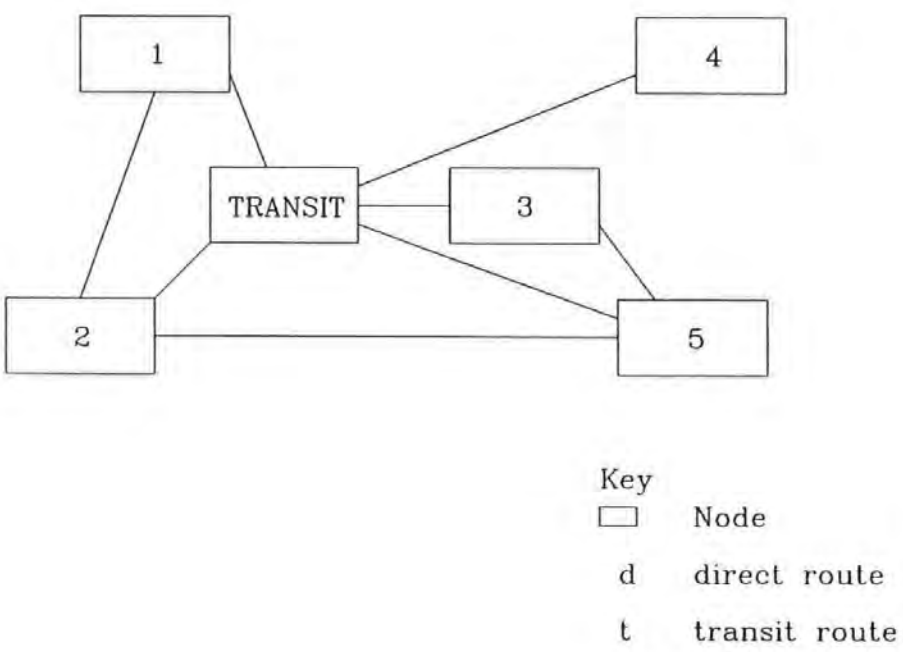


Figure 8 The Traffic Routeing Plan

Node	1	2	3	4	5	Transit
1	–	31				23
2	25	–			32	12
3			–		46	22
4				–		28
5		35	49		–	13
Transit	26	9	25	27	13	–
	Channels					

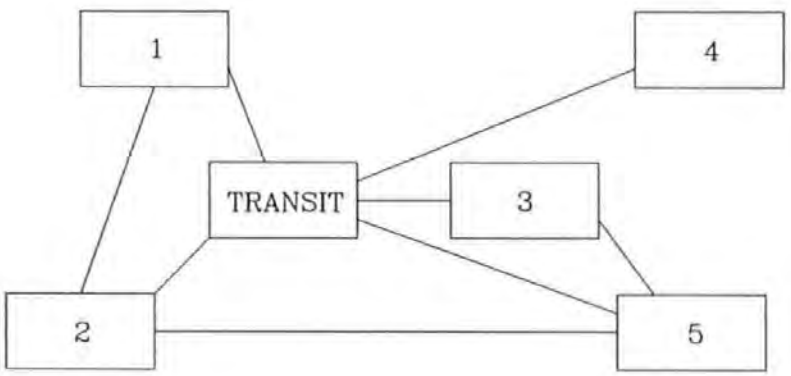


Figure 9 Dimensioning the Links

2.048 Mbit/sec link ports and total node traffic that equates to CPU size. Included in this final stage is the layout, dimension and costing of the transit link network.

Organisational issues, related to network designs, have a significant impact on costs or savings and are not usually ignored. For example, a plan that concentrates switching into large nodes could result in a more efficient maintenance organisation due to centralisation, better use of expensive support tools and sharing of overheads. However, this must be balanced against such a network being less resilient to equipment failure and traffic overload.

The traditional design approach makes for design studies to be conducted by computer using 'spread-sheets' representing the traffic and node location matrices; it is how most designers currently develop their network designs. Equipment node and link costs are presented in terms of discounted value of annual costs to take account of plant installation increments throughout the study period. For large network studies it is not be feasible to handle such a vast number of calculations by 'spread-sheet' methods, and the designers usually cost the network, in terms of annual charges, at 'snapshot' base dates, say at five, ten, fifteen and twenty years repeating the design procedure for each of the network scenarios to indicate the changes in network topology necessary to meet the changes in customer demand.

2.3.3 Design Accuracy

Network design is a complex activity: it must ensure that individual sub-network designs dovetail to produce a coherent development of the total network with the ability to cater for short-term unforeseen market requirements. It must also take

account of conflicting business pressures such as customer requirements for increased quality of service against capital constraints. It must have an accurate forecast that is adhered to within practical limitations.

However, this is not the norm for networks that have to be designed in practice. While it is possible to produce accurate forecasts for large areas, for example, the entire UK, they become progressively less precise as the area for the forecast is sub-divided down into streets and eventually individual potential customers.

Three possible courses of action that designers can take in response to forecast limitations are:

ignore the limitations;
provide more capacity than indicated by the forecast;
provide less capacity than indicated by the forecast.

The choice is influenced by the costs and means available to accommodate unforeseen demand in the network; for example, multiplexing on the access network.

Traditionally, a Telephone Company's (Telco's, e.g. BT), planning culture has been to avoid being short of capacity at all costs, and over-provision against forecast has been the natural response to any uncertainty. While this is obviously an effective way of reducing the chances of being short of capacity, it is not necessarily the most economic approach and leads to unacceptably high levels of node and link plant being installed even where the forecast proved to be correct.

Good network designers, when faced with an uncertain forecast, respond by balancing the risks and costs of taking corrective action for capacity unavailability against the cost of over provision. For example, consider two provision cost situations faced by a Telco showing:

marginal cost is high;

marginal cost is low;

where marginal cost is defined as the cost of providing an additional network element less the revenue accrued from its usage.

Against these consider two situations for corrective action in the event of capacity shortfall to meet demand:

cost of failure is high;

cost of failure is low;

where cost of failure is defined as the cost of providing additional network elements more than necessary to meet the forecasted demand.

These situations can be combined to form four possible scenarios:

i. high marginal cost, high cost of failure;

ii. high marginal cost, low cost of failure;

iii. low marginal cost, high cost of failure;

iv. low marginal cost, low cost of failure.

In cases i and iv there is a reasonable balance of cost between cost of prevention (over provision) and cost of failure (unplanned provision), and no adjustment to take account of possible forecast inaccuracy is likely to be the most economic course of action.

In case ii, there is no real benefit in over provision, there is probably a very high cost penalty, and an under provision rather than an over provision seems appropriate. Only in case iii can any financial justification be found for over provision.

While this example is simple, it illustrates that over provision is by no means the best choice in most cases. Thus a good network design methodology needs to take account of marginal costs and risks.

Also, for example, a telecommunication network is usually dimensioned to carry the BH traffic, which equates to peak demand during the day. This can be several times higher than the average demand over the day. The designer may decide that it is uneconomic to cater for the full BH traffic and instead only dimension the network for, say, 90% of the BH demand. Customers would then experience a severe deterioration in performance at peak periods if over-spill were not provided by another network. Whether this approach is feasible will depend upon the business environment and on the purpose of the network.

2.3.4 The Traffic Matrix

A telecommunication network consists of n nodes, distributed within a two-dimensional plane; each node i (i=1,2, ... n) is defined by its x and y co-ordinates with

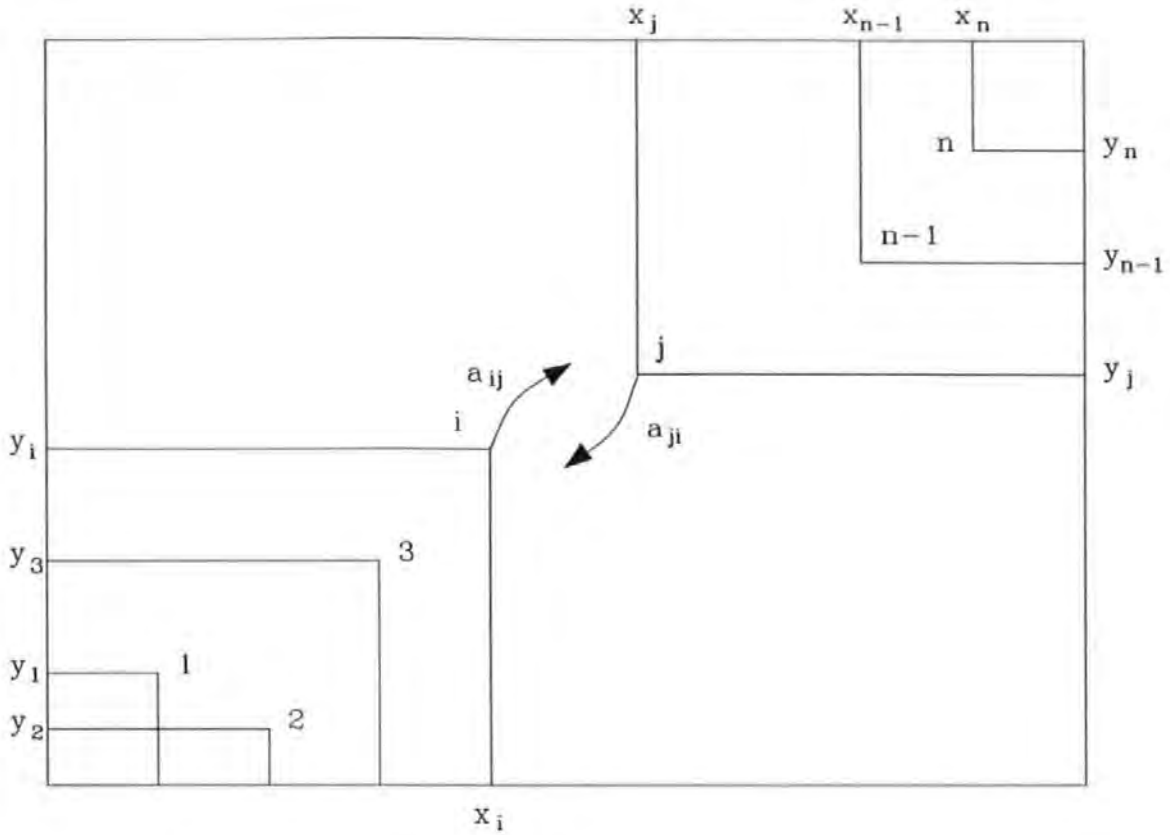
respect to some fixed frame of references (figure 10). These nodes represent traffic sources and sinks: each node could be a switch or computer terminal each wishing to communicate with each other, this communication takes the form of a traffic flow within the network. To describe this traffic, a traffic matrix $A = (a_{ij})$ is defined, in which A is an 'n by n' matrix where a_{ij} represents the traffic originating at node i and destined for node j. In the general case there may be more than one matrix since there may be more than one type of traffic, i.e. voice and data. The combination of co-ordinates (x,y) and traffic matrices A_1, A_2, A_3, ... completely describes the initial network.

2.4 Graph Theory & Telecommunication Networks

Graph structures are ideal for use in telecommunication network design when they represent nodes and links that form the network itself. The network graph is defined by a set of nodes and a set of edges, where each edge is a pair of nodes. As the edges are directed they are also called arcs. Arcs are represented by ordered pairs. The edges are often weighted with a cost, name or label representing the function or technology of the link.

2.4.1 Tractability

The optimising of large telecommunication networks is included in a family of problems that are regarded as being intractable. Such problems, though solvable in principle, are insolvable in practice when a large number of nodes are involved. When faced with the task of solving an intractable problem the only option available to the network designer is to solve a different, but related, and less difficult problem. The



$$(x,y) = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ & \\ & \\ x_n & y_n \end{bmatrix}$$

$$A = \begin{bmatrix} a_{11} & a_{1n} & \cdots\cdots & a_{1n} \\ a_{21} & & & \\ & & & \\ & & & \\ a_{n1} & & & a_{nn} \end{bmatrix}$$

Figure 10 Traffic Matrix Definition

relationship between the new and the old problem usually falls into one of two types: relaxing constraints on the outputs or restricting the number of inputs. Approximation, of either the inputs or outputs, is a typical method for relaxation. Approximation is what is happening when the travelling salesman problem is said to have been 'solved' by heuristic search. In reality, two different related problems are being confused: the problem of finding the best path, which is not being solved, and the problem of finding a 'reasonable' path, which is effectively solvable under various assumptions. **[3]**

Thus, in solving the best, or minimum cost, network design, the optimum may not have been found, rather a very close to optimum, or reasonable network has been achieved. This may not increase the cost by a large extent because of the normally flat minimum of such optimisation. **[4]**

2.4.2 Graph and Telecommunication Networks

Graphs look very much as telecommunication networks would be expected to look. The nodes represent sources and sinks of traffic, edges becoming links by which such traffic can be transmitted, paths between non-adjacent nodes becoming long distance communication routes and the weightings applied to each edge represent traffic capacities. The table below equates graph theory and network design terminologies

GRAPH THEORY	NETWORK THEORY
Node (Degree Zero)	Isolated Site
Node (Degree One)	Local Node
Node (Degree >1)	Transit Node
Edges & Arcs	Links
Weight & Flow	Traffic
Path	Route
Spanning Tree	Network
Star	Hierarchical-Star
Mesh	Hierarchical-Net

The basic objective of a telecommunication network design algorithm is to determine the cheapest way of connecting these n nodes. A typical problem and solution pair are shown in figure 11 and 12: they shows node traffic source and sink nodes. The nodes shown as circles represent switches, and squares represent, transits. Transits are those switching nodes that have been chosen to collect together traffic within a group area and then concentrate this traffic along a single route to other transits. This is much cheaper than connecting each node directly to every other node by way of high level links. The route for traffic from a node in group J to a node in groups K and L is as indicated. Transits are connected in an unrestricted way, whilst switches are connected directly to their closest transit.

The particular topologies relevant to telecommunication networks are known as hierarchical-star and hierarchical-net.

A hierarchical-star has the form shown in figure 13. A single node is chosen as the centre transit with n such choices available. From the remaining n-1 nodes, a further m secondary centres are selected; there are ${}^{n-1}C_m$ ways of achieving this.

A hierarchical-net has the form shown in figure 14. It is similar to a hierarchical-star in that there are two levels to the graph. In this case, however, the m secondary centres form any connected graph instead of being directly connected to the centre. The number of hierarchical-m-nets that can be constructed has been shown, by Grout **[5]**, to be $\binom{n}{m} m^{n-m} C_m$

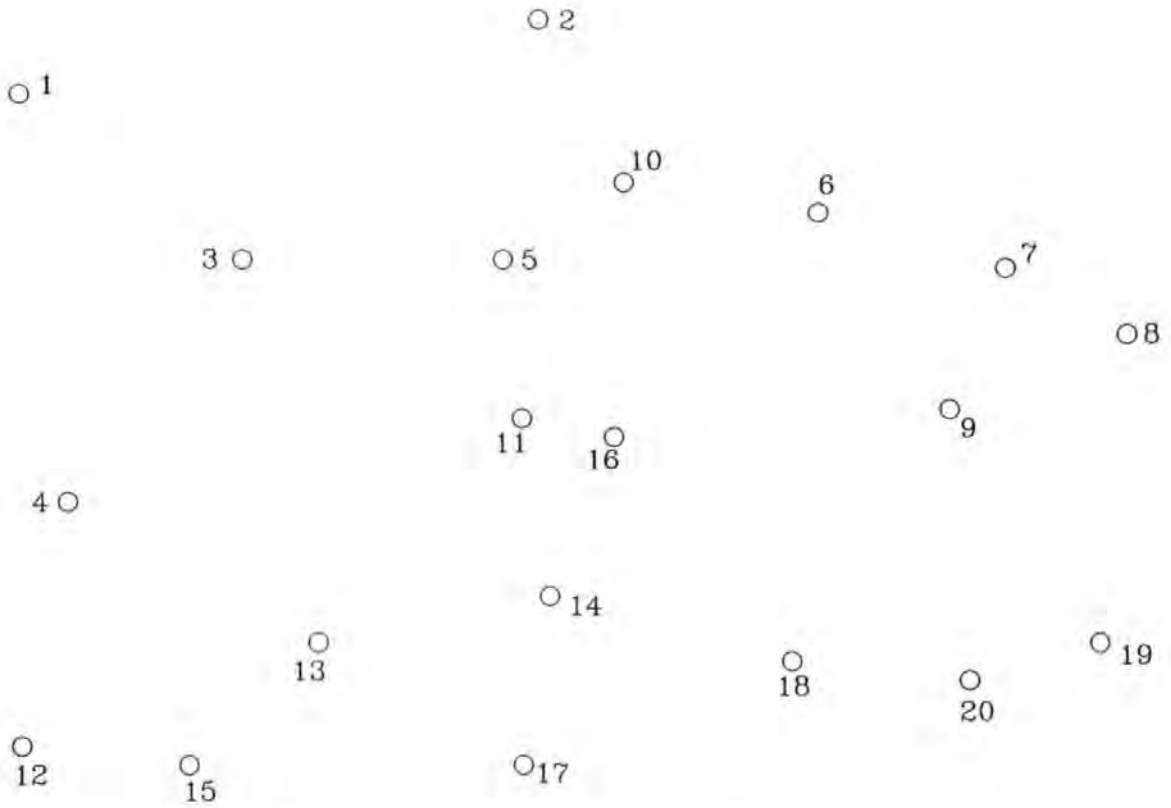


Figure 11 Typical Node Distribution

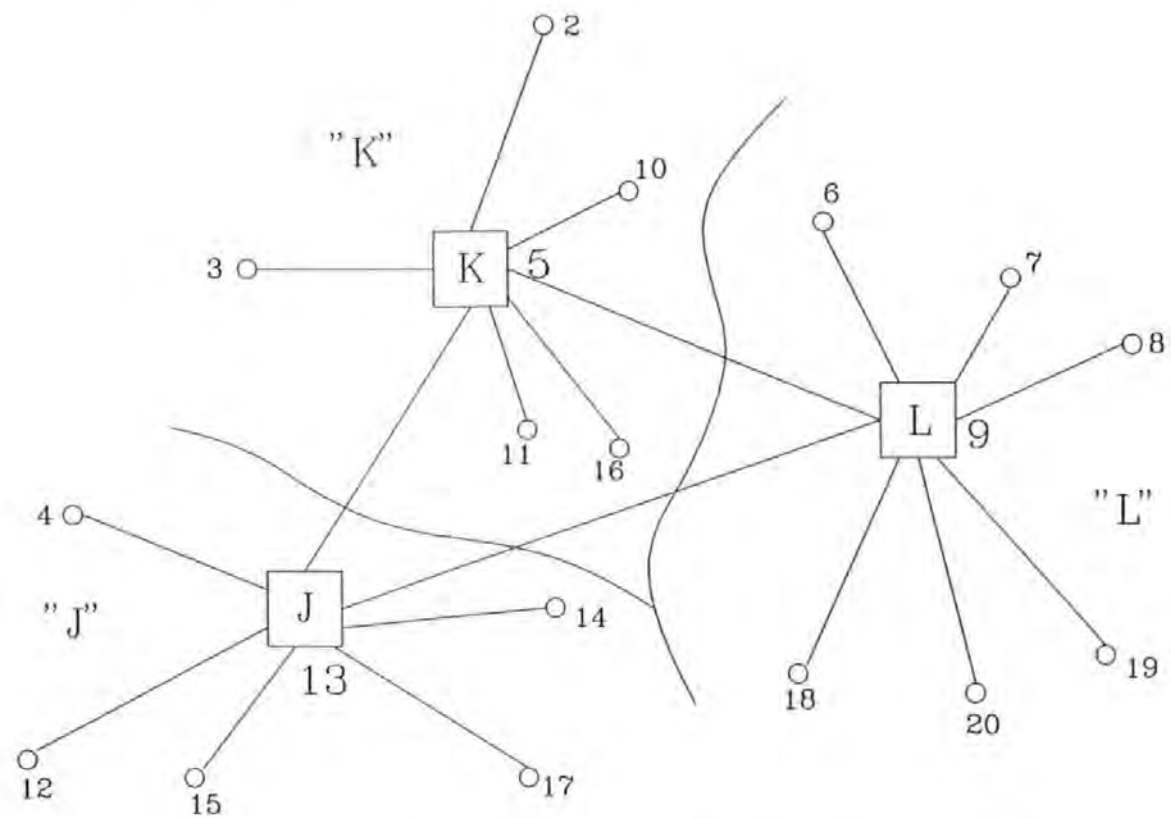


Figure 12 Typical Network Solution

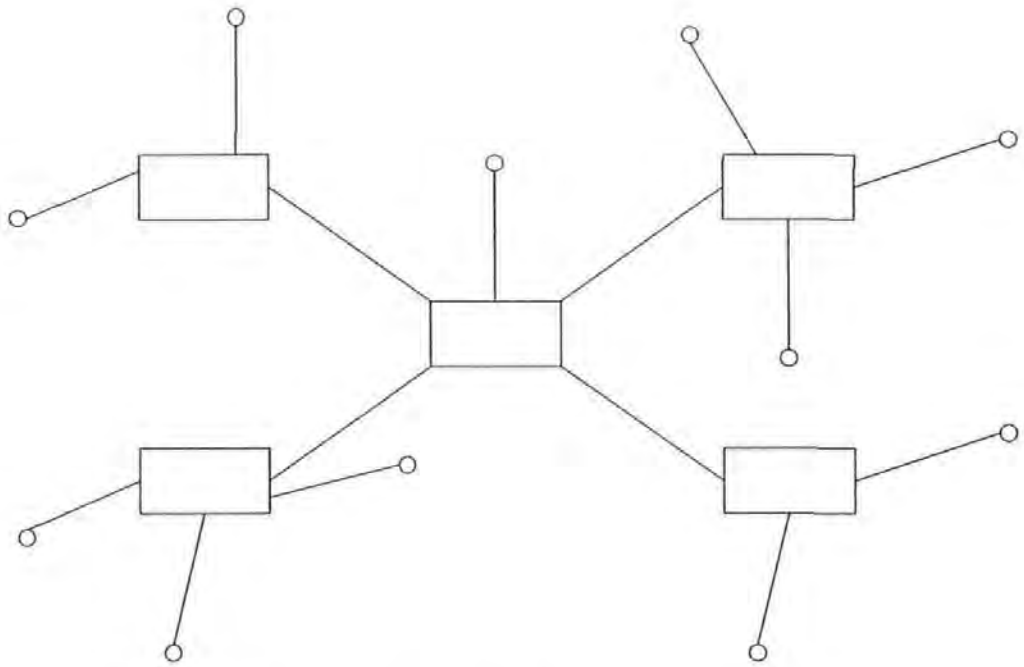


Figure 13 Network Hierarchical–Star

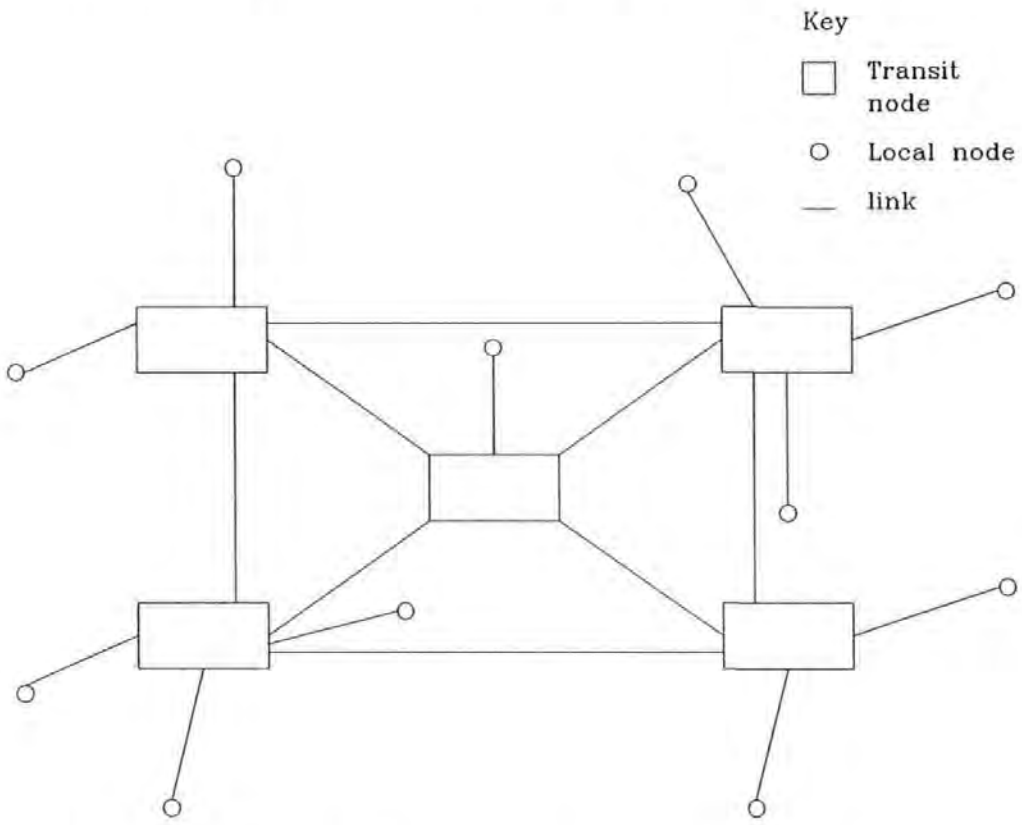


Figure 14 Network Hierarchical–Net

2.5 The General Network Optimisation Problem

Given the details of the initial network, along with the required grade of service (GOS) to be achieved and node and link costs, the objectives of any network optimisation algorithm are to determine:

the optimum number of nodes and transits;

the optimum position of each node and transit;

the optimum size of each node and transit;

which node is connected to each transit;

the optimum size of each node to transit link;

the optimum topology of the network;

the optimum size of each inter-transit link;

the optimum traffic routeing strategy within the core network;

the cost of the final solution.

2.5.1 Methods of Optimisation

There are numerous different approaches, each based upon graph theory, in which the general network optimisation problem is dealt with, most of which can be categorised into a smaller number of groups. One fundamental distinction to be made is between those methods that translate the problem into a more manageable one and those which work constantly on the actual network.

The most important dichotomy of all, however, is the classification of all methods into those which aim to give exact solutions all the time and those that only give

approximate ones. To provide an exact solution does not necessarily require that every possible solution network be tried out (in the same way that obtaining the solution of an algebraic equation does not require trying out every number in the equation). However, it can be accepted that even an efficient full optimisation process will be more involved and take more processing resources than a heuristic one. Exact methods may be suitable for smaller networks but as the number of nodes 'n' increases, however, the need for good heuristics grows stronger.

2.5.1.1 The Exhaustive Search Algorithm

An exhaustive search algorithm is one that systematically tries out every possible network before choosing a solution based on minimum cost. A brief description of such a method will indicate the probable difficulties that are likely to be meet in using this algorithm.

As with any exhaustive search method, the simplest way to proceed is to try out all possible solutions using a set of nested loops in which each variable of the network is controlled. The optimum number of transits is certainly a required part of the output and would logically constitute the outer loop within the main body of the algorithm. Within this group, the value of M (the number of transits) can be regarded as being fixed. If it is stipulated that a transit can only be placed at one of the existing nodes, then the problem of finding the optimum position of the transits reduces to finding the optimum combination of M transits from n nodes. This forms the next loop. Inside this loop, the position of the transits is established. The optimum size of each transit depends on which node is connected to each transit and the sizes of the links. The several different node-transit connection possibilities must, therefore, be tried out

within the third loop. With transit positions and node-transit links established, the general arrangement of the core network must be considered. Thus all the possible (valid) combinations of inter-transit links must be tried out within loop. The traffic routeing strategy within this core network can be considered as being taken care of when the topology of the core network is determined. The cost of the network must be calculated each time but is not part of the algorithm loop structure. As far as output is concerned, cost needs to be calculated once the final solution network is established.

Thus it is necessary to investigate every one of the possible combinations of M transits from n nodes. There are

$$^{n}C_{M} = \frac{n!}{M!(n-M)!}$$

of these. With the number of transits and their locations established, it is necessary to try out every possible set of node-transit connections. Each of the n nodes can be connected to one of M transits giving M choices. Therefore for n nodes, there are M^n calculations necessary.

2.5.1.2 The Add Algorithm

The Add algorithm was detailed by Bahl & Tang [6]. The operation of the Add algorithm begins with each node connected in a star formation to a central transit. All other nodes that could form transits are said to be closed (i.e. are not being used). Each node is then tried out, one at a time, as a transit and the one that produces the greatest improvement in cost is then opened permanently. This process is the continued until either all the transits are open or until no further improvement is to be obtained by opening another.

2.5.1.3 The Drop Algorithm

The Drop algorithm is very similar to the Add algorithm; indeed it was proposed along with the Add by Bahl & Tang [6]. It differs in that it begins with every node being a transit, i.e. open (instead of closed), and each is closed (instead of opened), one by one, until no improvement is possible.

2.5.1.4 The Tree Network Algorithm

A tree is effectively the most efficient way to connect a number of nodes. If the cost obtained was the absolute minimum then it was said that it was a minimal spanning tree. A similar situation exists in the practical situation. There are n-1 nodes, all of which are to be connected in some way to a single centre; they are not necessarily connected directly to the centre. It is acceptable in non-voice telecommunication networks to connect them to each other with only one node in each cluster being connected directly to the centre. Networks of this type are known as multidrop networks. Such networks are practical extensions of the ordinary spanning trees and, consequently, similar methods can be used to construct them.

2.5.1.5 The Chandy & Russell Algorithm

The branch and bound technique used for designing multi-drop networks, proposed by Little et alii [7], was developed by Chandy and Russell [8] to provide a logical and efficient tree generation strategy. It requires the sub-division of the solution space into successively smaller spaces. At each stage, the sub-space in which the required

solution exists is determined and the rest rejected. This sub-division continues until the optimum solution is found.

2.5.1.6 The Kruskal Algorithm

Kruskal's algorithm [9] is a constrained version of the branch and bound method. To implement this algorithm, one simply choose the cheapest link in turn, exactly as before, except that this time it is necessary to check to see that the constraints are not violated at each stage.

2.5.1.7 The Prim Algorithm

Prim's algorithm [10] is a variation of Kruskal's algorithm in which new links are not chosen from anywhere in the network but from the set of those links that are incident on the tree already formed. In this way, the tree 'spreads out' across the network.

As with Kruskal's algorithm, Prim's algorithm must be checked at each stage to ensure that the new traffic constraints are not violated.

2.5.1.8 The Esau-Williams Algorithm

The Esau-Williams algorithm [11] starts by connecting all nodes to the centre. In turn, the saving in cost achieved by disconnecting a node from the centre and connecting it to another node is considered (subject, of course, to the traffic constraints being satisfied). The replacement that leads to the greatest improvement is then effected permanently.

2.5.1.9 The Grout, Sanders & Stockel Algorithm

Grout, Sanders and Stockel's [12] approach to the problem of finding the minimum cost network is in two parts. The first is to replace the set of actual nodes with a reduced set of hypothetical nodes each representing a group of nodes. Fixed topology techniques are then employed upon this reduced set to find the location of the transits. Perturbation methods are then used based upon a 'look ahead' method of link removal to determining the transit interconnecting network.

2.6 Alternative Routeing Optimisation

The techniques for routeing optimisation have become more important in the UK as the process of deregulation offers the network designer a range of suppliers of links and network services.

Principles of teletraffic engineering [13] state, among other things, that the number of channels required to serve a given traffic (voice and non-voice) is a non-linear function of the amount of traffic; in other words, if a given amount of traffic requires a specific number of channels, doubling the traffic will not require double the number of channels for the same grade of service. Thus, it is often advantageous when designing to group together voice and non-voice traffic to take advantage of this effect. They also state when traffic between pairs of nodes is relatively small, a disproportional high number of channels are required for an acceptable grade of service. A compromise is to provide a small number of direct links as a first choice with calls that cannot find a free direct link being routed through a transit node.

Provided that there is sufficient traffic to keep the direct links highly occupied this will result in significant cost savings with little loss of efficiency.

In the majority of telecommunication networks, a call is allowed to take more than one route through the network. In such cases all routes are dimensioned to give the required quality-of-service. Traditional Poissonian statistics apply when calculating the grade of service. When re-routeing is allowed to overcome faults or overloading, Poissonian statistics are now more difficult to apply as traffic offered to the second and tertiary routes is no longer random. This later process requires a set of heuristic rules to calculate the resultant quality-of-service.

2.7 Limitations of Networking Algorithms

When designing a practical telecommunication network, it is necessary to determine the amount of equipment required in each part of the network that will satisfy a specific demand at a defined quality of service.

Whilst the above algorithms, based upon graph theory techniques, produce exact solutions, in practical environments there are almost an infinite number of solutions to the design problem, dependant upon the structure or technology points of view. The objective of the network designer is to determine, from all possible solutions, a reasonably efficient network as a starting point and to improve it by successive modifications in an attempt to achieve a minimum cost.

2.7.1 First Approximations of the Design

Cost optimal network designs made by traditional methods are unlikely to be truly optimal because subjective evaluations play an essential role in the decision making process. Moreover, several, sometimes conflicting objectives, are involved. Thus, the design now continues and becomes largely one of approximation.

Often, the network designer usually has to consider upgrading an existing network. It may be to the organisation's advantage, if new or additional transits be employed, to consider relocating the network nodes so that they are in an optimum position, i.e. sites at which line costs would be at a minimum. In a country the size of the UK however, where line lengths can never be excessive, the cost of relocation may be greater than any savings gained. Thus it is usual to upgrade a network based upon its present locations.

Practical constraints such as land availability for node buildings and obstacles also need to be considered; for example, rivers often form natural boundaries for traffic collecting areas.

The designer may also add a few links that intuitively seem beneficial. One quick way to enhance the design is by reduction of the diameter of the network (a measure relating to hops between nodes). The objective here is to minimise the total communication link distance, and therefore its cost; thus equalising the work-load over the available links and nodes.

Network designers will optimise routeing by a process called 'routeing by allocation'. This entails allocation of a certain fraction of the call traffic from each traffic source to each channel facility. This is preferred to differential overflow where attempts are made to place all source traffic onto one route, with the overflow going to another group and when that group overflows traffic goes to another group and so on. The modelling of such a process requires a great deal of conventional computing power.

2.7.2 Other Design Problems

Other problems met by the designer in considering the practical networks are the subjective, un-quantifiable and practical constraints. The contribution of the network designer now becomes one of balancing the quantifiable with the un-quantifiable. For example, the following engineering criterion will have their effect.

Costs:

for different types of cables per unit length;

for nodal equipment;

for sites and buildings.

Topology:

signalling and link limits;

traffic losses (Grade of Service).

Present state of network:

existing nodes;

existing duct systems;

customers access cables;

transit link network.

A practical network design needs to be such that a high degree of flexibility is available. However, this will have a cost penalty: the longer the design period the greater the uncertainty and the shorter the design period the greater the overall cost. The approaches available to the designer include over or under dimensioning. Over dimensioning has the risk that equipment will be used either in the long term or never. Under dimensioning has the risk of requiring additional and, in some cases, ad-hoc installations resulting in an unnecessary complex network topology.

When dealing with the flexibility of the network the designer must additionally:

consider a series of cost parameters such as: recovery values of equipment, removal costs, equipment lives and cost of rearrangements;

perform sensitivity analysis on each new alternative considered aimed at identifying the important factors. This can be achieved by the changing of the traffic matrix and/or costs;

analyse and scrutinise the most costly items of network hardware;

aim at a homogeneous solution that is likely to be more flexible than the optimum network based upon peculiarities only found in that particular case. The designer should verify, before adopting the solution, that the peculiarities are firmly based.

2.8 Quality of Service

Quality of service will also have its impact on the eventual design. It covers the areas of traffic grade of service, cost of failure and network security. These ideas are difficult to describe in mathematical terms.

2.8.1 Traffic Constraints

Given a measure of traffic flow in erlangs, (one erlang of traffic is defined as one channel occupied continuously for one hour), it is necessary to find the number of channels needed to carry it. Under provision of channels will result in calls being lost and this will happen when all α channels are in use and a $(\alpha+1)$th call is attempted.

Network grade of service is the probability of a call being blocked some where in the network before it has been routed to the destination customer. To find the network grade of service, it is necessary to look at the cumulative effect of links and nodes both in series and parallel. The most widely used formulae for evaluating network grade of service are those developed for evaluating the blocking probability in switches. Lee [14] developed the basic principles of the mathematics for calculating congestion. He proposed the transformation of switching matrices into linear graphs where nodes represent switching stages and edges the links, the linear graph representing all the

paths from a particular inlet to a particular outlet. However, with a few simplifying assumptions the theory is readily adaptable to an approximate approach to calculating network grade of service.

If individual portions of a network are assumed to be independent (i.e. the supply network is independent of the core network) then the overall blocking probability can be written:

$$B = 1 - [(1 - P1)(1 - P2)(1 - P3)(1 - P4)]$$

where P1 and P2 are the probabilities of blocking on the supply network and P3 and P4 are similar probabilities on the core network.

As individual links are usually designed with a grade of service loss probability of 1 call in 100 (P=0.01) then B approximates to

$$B = P1 + P2 + P3 + P4,$$

expressing the rule that losses are additive.

It is still necessary to calculate P3 and P4, the trunk network probabilities. In the case of a star network P3 and P4 are simply in series with the access network and hence are additive to P1 and P2. However, in the case of a mesh network, with its non-series parallel core, it represents a computational difficulty problem to solve, as highlighted by Lee **[14]**. It is now necessary to resort to simplifying algorithms to

give 'indicative' blocking probabilities for the trunk network and a rule that losses are additive will give a worst-case resultant.

The network designed so far has been to a fixed grade of service principle: that is, a grade of service is fixed for the network as a whole and the component parts engineered or dimensioned accordingly.

This approach has two practical defects.

The first is that grade of service takes no account of the fact that large groups of channels or nodes are not only more efficient than small groups but also more sensitive to congestion and faults, i.e. their grade of service deteriorates by a higher proportion for a given percentage increase in traffic. In practice, therefore, it is more effective to engineer different stages of the network for different grades of service rather than a single end to end criteria.

The second is that a fixed grade of service principle does not, in its self, specify what the grade of service should be. Grade of service design values are chosen to give a level of capacity that are necessary to give a workable quality of service for a reasonable price: quality now becoming the subjective test of network performance. A practical solution is that the optimal number of trunks, i.e. a combination of channel and nodal ports involved in a call, is such that the revenue from an additional trunk is equal to the cost of providing it, i.e. zero marginal cost. Each item of network equipment is now installed at the level at which the value of marginal increase in traffic equals its marginal cost. Thus the design grade of service is driven by the objectives of the business.

Value, to the service provider, must be attached to the cost of losing calls. Unless congestion is very severe, calls are rarely lost for ever, i.e. abandoned, but repeated until successful. If all lost calls were repeated until successful there would be no loss of revenue resulting from a poor grade of service.

Cost to the customer can be identified as:

time spent redialling;
time wasted between attempts to redial;
delay in completing a call;
frustration and annoyance generated by congestion.

A reduction in congestion can be viewed as a reduction in cost for which the customer will pay an increased price! From a business viewpoint, concern centres upon the proportional of call attempts resulting in number engaged or no reply as this is ties up equipment and could result in loss of revenue.

2.8.2 Network Quality

The quality of the network, in terms of probability of call failures, is automatically accounted for in the network design process. In summary, the grade of service is expressed as the percentage of calls lost or the proportion of time during which all the channels on a route are engaged. It is possible to enhance the grade of service by taking account of the cost of failure **[15]**. With private networks, it is possible to provide a grade of service significantly different from that offered in the public network, but at a correspondingly different price. Indeed, some private networks are

provided on the basis that they are cheaper, but at a worst grade of service, than an equivalent service from the public network.

2.8.3 Network Reliability

Network reliability is affected by two basic types of uncertainty that may occur in the development of the traffic matrix, these will affect the optimum network design.

Firstly, traffic distribution develops as forecasted but there is a change in timing: if the demand appears earlier than expected the effect is not too significant from an economic point of view but the demand appears late it results in poor use of the capital investment. Secondly, if demand does not happen in the life of the equipment, or the traffic demand was forecast in one place but appears in another. This is one of the worst cases, as it is equivalent to an underestimate in one place and overestimate in another at the same time.

It is unlikely that the designer will have a totally 'green field' situation, there is usually a need to migrate from the existing to the new whilst maintaining network grade of service but also minimising transitional cost.

2.8.4 Network Security

A number of design changes are necessary to take account of network security, i.e. protection against network component failure:

When the traffic between two points is greater than 100 erlangs, a common practice is to carry the traffic on different physical paths. If there is a fault on one of the paths, at least part of the traffic can be handled on the other. It is common to limit the number of customer's lines in one building to 60,000. In a similar way, the maximum number of erlangs handled by the transit node is limited to 16,000, i.e. 22,000 channels at 0.7 erlangs per connection. Back-up equipment can also be provided, e.g. the provision of at least two physical paths to each pair of transits.

All these constraints lead to a network structure that differs from the optimum.

Two design approaches can be used to enhance the reliability and security of the network: firstly determine the economic penalties caused by constraints imposed for reasons of security. This design activity is one whereby the cost differential between the optimum and reliable networks is identified. Secondly, design the network with account taken of the probability of a catastrophic event occurring, i.e. loss of transit or main link. This design process is one of enhancing the capacities based upon the probabilities.

2.9 Implementation Process

Whilst not explicit in the network design, it will be necessary to phase the implementation of the network. The following rules can be used to minimise the cost of implementation:

exclusively 'City Core' and 'Suburban' development of the network should be avoided;

minimum network coverage of 'Fringe' areas should be obtained at the initial implementation stages.

2.10 Uncertainty

The automation of the telecommunication network design process to produce practical network designs poses a unique problem that needs to be overcome: that of uncertainty.

Facts as 'the switch is digital' can be replaced by possibilities 'it looks as if it may be a digital switch' or probabilities 'the chance that it is a digital switch is 70%'. Coping with possibilities or probabilities means introducing an approximate reasoning system. The problem of coping is compounded by the need to reason with facts that are only opinions or beliefs of the probability or possibility of an event having occurred: for a review, see Caudill **[16]**.

Uncertainty has significant effects upon the eventual design.

2.10.1 Uncertainty in Telecommunication Network Design

There are two fundamental areas were uncertainty impacts upon the integrity of a telecommunication network design.

The first is that, although the ideal solution changes with time, the initial design has to be based around a single time, normally at the beginning of the design period. Thus it is essential to allow for the uncertainty in the accuracy and values of

predicted data, particularly identifying the most sensitive inputs so that either different or more resilient designs can be drafted. It needs to make allowances for network growth, changes in technology and compensate for deviations from the original design data, e.g. demand, prices, availability of resource and hardware, etc. The uncertainty of this data must be taken into consideration, with particular attention being paid to the identification of the most sensitive inputs so that either different or more resilient designs can be drafted.

A second, and more fundamental area, is the uncertainty that resides at the initial design stage. Here it is necessary to generate a 'source-to-destination' traffic matrix. The measurement, estimation or allocation of traffic values forms the base data to all network design models. It is at this stage where uncertainty, in both data and application, will have its greatest effect, as errors made at this stage are promulgated throughout the network plan.

2.11 Computer Aided Telecommunication Network Design

Telecommunication network design consists of the optimisation of capital investment and dimensioning equipment for the area covered by the customer community, meeting objectives set by the eventual owners of the network. The prime objective of a design is to provide the right equipment at the right place at the right cost to satisfy expected demand and give an acceptable quality-of-service. The complexity of the design process has resulted in a growth in numbers of computerised optimisation computer packages [5]. Unfortunately, such packages lose sight of the practical problems that need solutions. Here, constraints of equipment availability, the uniqueness of the characteristics of the community served by the network and the

inability to be fully cognisant of all the data necessary for the design process remove such packages far from the pragmatic environment in which their authors wish them to be employed.

In such circumstances, heuristics are necessary and practical restraints of the component parts of a network need to be incorporated, then an expert system approach proves to be a more useful tool to the network designer.

Kai-li **[17]** categorises all methods of optimising procedures used in the network designing process into two distinct groups.

The first group is classified as the *expert assist* set of models. These models do not produce optimised solutions themselves, but compute or simulate the consequences of alternatives under the control of skilled network designers. The designer, using experience or intuition, proposes a series of different designs or designs, inputs them into the computer, then evaluates and compares the results. For a good review of assist models see Roth et alii **[18]** and Mantelman **[19]**, whilst Ferguson et alii **[20]** gives details of an assist model used by Bell Canada.

The second group is classified as the *logical search* set of models. These models create a 'search space' that contains possible solutions to the problem, the computer then performs a systematic logical search by using theorems and heuristic algorithms to find the optimal solution. Because this group of models is based on well established algorithms, they should give the optimal (or close too optimal) result: see Grout **[5]** for a review.

In practice, however, when confronted with real network designing problems, it is not always possible to match the required theoretical conditions required by the various algorithms. For those that can, usually many assumptions and simplifications need to be made, causing the model to be quite different from the problem, thus invalidating the optimal result.

Tange **[21]**, argues that it is not possible simultaneously to optimise all variables in the network designing process. Iterative methods, providing a number of sub-optima, are normally used giving an acceptable approach to an optimal network plan. Gupta **[22]** has suggested that network designing is more of an art than a science! He argues that it is often impossible to describe factors such as accommodation, regulation and organisational issues as a mathematical model, they can be better simulated by an expert system.

Furthermore, the optimisation of telecommunication networks is included in a family of problems that are regarded as being intractable. Such problems, though solvable in principle, are insolvable in practice when a large number of nodes 'n' are involved. Penrose **[23]** states *problems in (as these) are regarded as intractable for a reasonably large 'n', no matter what increases in operational computer speed, of any foreseeable kind, are envisaged. The actual time that would be taken* (to solve the problem) *rapidly becomes longer than the age of the universe.* Thus the network designing problem, with its large number of nodes, has no efficient logical search solution. It is usually necessary to make assumptions about either the input or output of the large problems and as such the use of heuristic methods of search and problem solution can have equal status to quantitative methods.

2.11.1 Expert Systems for Telecommunication Network Design

The introduction of natural computer languages with advanced micro-electronics conceived the birth of expert systems in the late 1970s. By 1980, Japan had announced that a new, fifth generation, computer system was being developed that would revolutionise the way computers operated, opening the door to Robotics and Expert Systems.

In 1986, the start of this research programme, *using expert systems to solve telecommunication problems is still a new and fairly sparse area* **[19]**. The Financial Times **[24]** stated, among other things, *the commercialisation of expert systems is overdue, hundreds of expert systems (in the USA) have been created over recent years but the vast majority are either research projects or at best commercial prototypes.*

During the late 1980s, *the application of AI to telecommunication network design was well understood but implementation is slow and cautious* **[25]** and by 1991 few expert systems have been developed; notably those of NTT, the Japanese Telco **[26]**, and the US Bellcore **[27]**; neither of the expert systems are true design tools, rather they are aids to the network design process. Throughout this period, uncertainty modelling in expert systems remained ad-hoc: for example, the 1990 IEE Colloquium on Reasoning-Under-Uncertainty **[28]** covered symbolic and numerical methods; probabilistic reasoning; belief modelling; Bayesian and Fuzzy Logic. They concluded that no single uncertainty calculus was ideal.

Few attempts have been made to use expert system techniques for the design of telecommunication networks. An overview of design areas addressed is given by Lees **[29]**. Salasoo **[30]** describes a Bellcore system that supports the design process by integrating knowledge from several databases and user models to guide the design engineer. Solo et alii **[31]** have developed an investment tool that combines technical knowledge about network structures with managers subjective assessments so that the network evolution can be examined for differing assumptions regarding demand and facilities. Kamimura et alii **[26]** describes two NTT models that can be used for the comparison of public network services over private, in terms of cost, neither model takes a holistic approach to the design process.

An expert system for telecommunication network design should give designers the capability to:

know and analyse call traffic;
select the most effective link vendors;
compute tariff costs;
optimise the choice of links;
automate teletraffic calculations;
design a least cost network that meets customers demands.

Expert systems are unlikely to provide a universal solution to the network designing problem by themselves as a large portion of the problem still requires a quantitative and analytical approach. This is especially true if the overall problem has been divided into sub-problems. The sub-problems can often be formulated into mathematical models, a combination of heuristic and traditional programs should

provide the best and most efficient solutions under the overall management of an expert system. The expert system breaks down a large network problem into its component parts, uses different methods to solve each of them, and then combines these results to create an overall solution.

Thus, it is necessary to divide and sub-divide the overall problem into smaller problems until some parts become mathematically tractable. An appropriate degree of coupling between the sub-problems has to be maintained, however, if the solutions for the sub-problems are to stay near-optimal for the global problem. In this context, the process of sub-dividing can be viewed as a heuristic approach to the overall network optimisation problem.

With an expert system approach, it is therefore necessary to settle for near-optimum designs obtained by interactively solving the series of sub-divided problems. This should not be cause for concern as the cost minimum is fairly flat, slight deviations from the optimum point are normally of little importance **[4]**. The mechanism of sub-division is fundamental to the success of the process and will be problem dependent.

CHAPTER 3

EXPERT SYSTEMS and UNCERTAINTY

3.1 Introduction

This chapter introduces the philosophy of intelligence and follows an argument through to a definition of an expert system. The fundamental properties of expert systems are then discussed and how they handle uncertain information is investigated.

3.2 Expert Systems

As long ago as the 17th century, Descartes [32] looked at the fundamental truths of knowledge and intelligence. Russell [33 & 34] later developed these theories for a more complete understanding. Both Descartes and Russell questioned the scientific method of moving from hypothesis to theory and law. Ayer, however, maintained that *all inductive reasoning is illegitimate and there are no well established theories* [35]. What makes the application of philosophy to the scientific process problematic is that *it prevents one from discriminating between stronger and weaker evidence and possible superior and inferior options* [36].

When dealing with uncertainty, Popper states [37] that is is impossible to ever *prove any theory is correct (given a finite amount of evidence).* The Reverend Thomas Bayes evolved a non-monotonic probability theory that attempts to overcome this problem [38]. Bayesian inference examines all the evidence that is relevant in making a probability assessment of truth. Aleluias [39] identified a defect in Bayesian

inference and says that *its naive application engenders a hunger for probabilities, which can rarely be satisfied, and a surfeit of computation.*

The first reference to a machine with intelligence is found in 'The Politics' [40]. Here, Aristotle writes: *proper tools will be essential for the performance of a task ... that every tool we had could perform its task ... itself perceiving the need ... of their own accord.* Two thousand years later, Lady Lovelace recalling Babbage's Analytical Engine said *it has no pretensions to originate anything. It can (only) do whatever we know how to order it to perform* [41]. The twentieth century development of cybernetics [42] lead to debates on whether a machine could possess intelligence. The British Medical Association concluded *not until a machine can write a sonnet ... because of thoughts and emotions felt ... could we agree that a machine equals brain that-is ... no machine could feel pleasure at its successes ... be charmed by sex ... or be depressed when it cannot get what it wants* [44]. Turing [45] proposed a pragmatic view of machine intelligence and asks *what will happen when a machine takes the place of a man in a game (hidden from view) will the interrogator be unable to identify the machine?* Such is the development of expert machines or systems, that today no one has passed Turing's test [46] although a whole new branch of computer science called artificial intelligence has developed from it [47].

The following definitions are based upon an excerpt from Kehoe **[24]**: *an expert system comprises two major parts, a knowledge base and inference engine.*

The knowledge base contains all the information relating to a particular application ... it is programmed with information obtained from experts and text books in a given field.

The inference engine contains a set of rules that should be applied to the knowledge base to reach a conclusion.

When a question is asked of the computer 'expert', it searches through its knowledge base, applying rules in a what if/then type procedure until it comes up with an answer. The process of developing an expert system involves extracting knowledge from human experts and coding it into computer language.

A modified Feigenbaum **[48]** definition of an expert system as *intelligent computer programmes that use knowledge, and inference procedures, to solve problems that are difficult enough to require significant human expertise for their solution. The knowledge necessary to perform at such a level, plus the inference procedures used, can be thought of as a model of the expertise of the best practitioners in the field.*

Knowledge of an expert system consists of facts, rules, data and heuristics. The facts, rules and data constitute a body of information that is widely shared, publicly available and generally agreed upon by experts. The heuristics are mostly private, rules of good judgement or thumb and intelligent guessing that characterises expert level decision making in the field.

Expert systems view heuristic knowledge, (the way knowledge, experience and intuition work together to solve a problem), to be equally important with factual knowledge, indeed, to be the essence of what is called expertise

3.3 Advantages of Expert Systems for Network Design

When used for telecommunication network design, expert systems have some advantages over conventional computer systems.

Conventional computer systems can only provide help on those problems that have a mathematical or statistical core and are of a routine data processing nature. Parts of the telecommunication network design problem do not fit into this category; synchronisation and demand assessment for example. Instead of requiring elaborate calculations, they involve inferential reasoning using a large professional knowledge base and the application of a set of heuristic rules relevant to the problem domain. Also, the complexity of the problem together with a large number of possible variables makes the design problem difficult to describe mathematically. Expert systems are specifically designed to tackle this set of problems since their whole focus is on knowledge. Expert systems are divided, not into code and data like traditional programs, but into a corpus of knowledge and a comparatively simple mechanism for applying that knowledge in a way to solve problems.

The power of expert systems does not come principally from its 'inference engine' but from the richness, pertinence and redundancy of the knowledge itself.

Conventional programming techniques are inadequate for these domains because they tend to be deterministic and possess little or no redundancy, i.e. for any given input, there is a single computational path that is always followed and a single mechanism capable of producing the correct output for that input. This means that, although very compact, these programs have no facilities for coping 'intelligently' with circumstances unforeseen by the designer.

Conventional techniques are also individualistic and hence difficult to modify. They lack the human self-awareness of the techniques they are using to solve a problem and they are unable to reason about or explain their own mechanisms. This means that they are useless either as a method of codifying expertise or as a vehicle for communicating it.

In contrast to conventional systems, expert systems are more flexible in their decision making; for example, they can incorporate information of any kind of heuristic that may be relevant if something unexpected happens. Their ability to handle qualitative judgments also means that they can weight facts and assumptions and, in a sense, choose the most appropriate decision strategy for each problem presented to them.

Expert systems enable the application of computer power to many problem areas in the telecommunication industry where analytic humans are successful because of their professional knowledge and experience and their use of an appropriate set of heuristics for solving the problem.

Expert systems have one major disadvantage in comparison with human experts; they have a highly specialised by extremely narrow domain of intelligence. Outside this domain they know nothing. In other words, expert systems lack any general world knowledge or common sense. In some situations, this lack of general knowledge could seriously affect the decision the system makes.

3.4 Introduction to Modelling Uncertainty

There are currently four generic techniques for dealing with uncertainty in expert systems. Each takes one of the following forms:

linguistical;

logical;

statistical;

ad-hoc.

Linguistical techniques advocate the use of language to describe uncertainty thus allowing a calculus to mimic the expert. Utilisation of such techniques in an expert system requires the conversion of the terms into numbers for their subsequent manipulation. Unfortunately, there is considerable variation in the way different people interpret phrases and this interpretation is context dependant.

Logical techniques that use traditional logic have an inability to handle conflicting knowledge. This is due to its intrinsic monotonic nature. Fuzzy Logic, with its truth values ranging between zero and unity, overcomes this problem. However, in application it suffers the same problems that exist as for linguistic techniques.

Statistical techniques use a system of weights associated with data based upon the users' interpretation of the frequency with which the data would be true in a long series of trials. The Bayesian approach is a recursive technique where the posterior probability becomes the new a-priori probability of data when it is reused. Probabilities tend to be misused as it is often not possible in practice to determine the probability of an event, rather, a personal probability is specified. Different people can and do assign, quite legitimately, different probabilities to the same event.

Ad-Hoc techniques have been developed to solve particular problems. Whilst they work well for that particular argument many expert systems reuse the techniques

even though the mathematical premise of their use under such circumstances can be questioned.

Current methods of dealing with uncertainty in expert systems will be shown to be inadequate for telecommunication network design; their simplicity coupled with a lack of knowledge of their compilation means the intrinsic richness of uncertainty values to add to the design process has been lost.

A common significant failing of current techniques is that they are singular in dimension. They do not take account of the different attributes of uncertainty that need to be propagated in different ways through the expert system. Indeed, some attributes of uncertainty may be regarded as mutually inclusive whilst others are mutually exclusive. The composition of the resultant uncertainty contains intrinsic richness on the problem being solved, for example, those rules or data that contribute greatest to the 'doubt' of the final design can be found if the resultant uncertainty can be decomposed.

3.5 Expert Systems and Dealing with Uncertainty

The term 'uncertainty' has been given a wide interpretation within expert systems and appears to be used whenever reasoning by strict logical implication is not considered possible; this may be due to inadequacies in the knowledge-base, insufficient or unreliable data on the subject of interest, or because of stochastic relationships between propositions.

Uncertainty is a capacious term, used to cover a large number of concepts. It may arise because of:

incomplete information - what will be the regulatory environment?

disagreement between information sources - will traffic levels decrease in times of recession?

linguistic imprecision - what is meant by a significant increase in traffic?

variability - what is the calling rate per extension?

Uncertainty may be about QUANTITY, e.g. the slope of a traffic forecasting curve or about STRUCTURE, e.g. the shape of the traffic forecasting curve.

It is even possible to be uncertain about how uncertain one is!

The variety of types, sources and a lack of detail differentiation of the form of uncertainty generates considerable confusion. This thesis considers it important to distinguish clearly between the different types and sources of uncertainty since their treatment and promulgation through the expert system needs to be treated in different ways.

3.5.1 Overview of Current Methods

Evidence is indicative: it is whatever makes clear the truth or falsehood of something. It is necessary to describe methods for combining evidence derived from different premises or from different rules.

The association between premises and conclusions is defined by the weight with which the truth of premises supports the truth of conclusions. A difference exists between a monotonic and a non-monotonic association; in that the former changes the degree of the conclusion to a specific value, while the latter changes the degree of the

conclusion proportionately according to the weight of the association and the certainty of the facts. Indeed, a monotonic relationship requires that the addition of new information does not invalidate previous information, whilst non-monotonic relationship allows complete changes from true too false, by the addition of new information.

Association take one of the following forms:

antecedent ⇒ consequences

e.g.) IF x is System X THEN x is digital;

premise ⇒ conclusion

e.g.) IF accuracy is 10^{11} THEN clock is atomic;

situation ⇒ action

e.g.) IF stability is 1 slip per day THEN turn off clock.

With such associations, when the conditions are true, then the conclusions are also true. However, a fusion or 'combining' technique of conditional data before the generation of a conclusion is required. This is particularly true if an expert system produces a conclusion by different routes. Thus the use of some form of weighting factor in the combination process is necessary. Such weightings would take account both of data-relevance and, perhaps, statistical dependence.

Indeed, a weighting function needs to be taken into account whenever data is not categorical, i.e. true or false; here a degree of uncertainty exists as is the case in

expert behaviour that is full of heuristics when facts may not always be true or at least 100% true!

Clark **[49]** shows that uncertainty can arise in relation to the applicability of a general heuristic or how best to deal with lack of knowledge: *Information concerning the type of uncertainty is lost since sources of uncertainty are indistinguishable once compiled as probabilities ... ignores much information that is necessary in successful reasoning about uncertainty. A related point is that once knowledge has been compiled as probabilities it can not be employed in other tasks.*

Weighting evidence to resolve conflict may be necessary or desirable, but it is not the obvious method to use: indeed, it may be less useful than numerous least commitment approaches, such as hedging, worst-case analysis, taking the central value, deciding not to decide and so on **[50]**.

Despite what authors claim, there does not seem to be one calculus that is 'the best' for all situations.

Also, O'Leary et alii **[51]** show that confidence factors, when used in an expert system, can be input incorrectly for a number of reasons.

> Firstly, the wrong confidence value could have been recorded or keyed into the system.
>
> Secondly, the expert is likely to satisfy rather than optimise, as a result the quality of their estimates is just good enough to meet expectations.
>
> Thirdly, an expert may find it difficult to express confidence or belief as a number.

Fourthly, the expert does not want to be replaced by an expert system.

Current formalism provides no method for combining non-independent evidence other than omitting it in the same rule.

How can uncertainty be modelled?

A review of the many domains that use non-categorical knowledge including Market Research **[52]**, Medical Diagnosis **[53]**, Engineering **[54]** and more general surveys **[55 to 60]** have shown that, in essence, current modelling of uncertainty takes one of the following forms:

linguistic;
categoric (logic);
probabilistic;
ad-hoc certainty factors;
Bayesian logic;
fuzzy logic.

Henkind et alii **[61]** analyses three methods of dealing with uncertainty, Bayesian, fuzzy and ad-hoc, and shows the application of the theory to be inadequate.

3.5.2 Existing Methods

3.5.2.1 Linguistic Modelling

Fox [62] and Averkin et alii [63] advocate the use of linguistic terms to describe uncertainty. They argue that it allows a calculus to mimic better the expert and give a more natural explanation.

Fox believes that people *can, and usually do, deal with uncertainty in a similar way.*

As an illustration, consider the analogous distinctions when dealing with place and time. 'Place' can be represented by means of x and y co-ordinates; but in problem solving it may be more convenient to distinguish certain important places like 'here', 'somewhere', 'everywhere', and so on. Similarly, time can be represented by reference to a clock, but it is sometimes more useful to distinguish 'now', 'sometime' and 'never'. Although the quantitative meaning of such terms is ambiguous, their logical intention is not and often there is no need to commit oneself to a quantitative interpretation to make use of them.

Fox argues that we freely employ logical distinctions about belief and expectations without having to commit to a specific quantitative interpretation. He defines the following uncertain logical distinctions.

'X-is-possible' *if no conditions that are necessary for X are violated. X may have no necessary conditions, or we may be ignorant of the state of these conditions; these circumstances do not affect a statement of possibility. There may also be evidence for or against X without affecting possibility.*

*'**X-is-plausible**' if X is possible, and there is an argument in support of X or the arguments for X are stronger than those against (however, the term "stronger" is defined). In the case where arguments exist for and against X (e.g. from different assumptions) then use can be made of strategies to decide plausibility. These include preferring arguments that invoke general principles rather than ad-hoc special cases, or those that are short to those that whether or not there is evidence to support X and even if there is evidence against (e.g. if argued that the evidence is compromised).*

*'**X-is-probable**' if X is possible and there is at least one item of evidence in favour of X. When evidence goes both ways, we may again choose various strategies for resolving the conflicting preference to direct evidence over indirect.*

*'**X-is-certain**' when a sufficient condition for X is true. The sufficient condition for X may be evidence (e.g. the evidence of one's eyes) or an argument (e.g. an argument for exclusion of alternatives).*

Fox defines other closely related terms, including negative terms (**not possible, not probable**), contrast terms (**impossible, implausible**), and complex terms (**inconsistent, ambiguous**).

Additional uncertainty terms are defined by Fox that are vague in the sense that they do not make a commitment to the particular type of uncertainty they depend upon. For example:

*'**X-is-believed**' if it is reported by a credible informant or a reliable device or procedure, or if it is the most probable or plausible of the alternatives.*

*'**X-is-likely**' if the summary of evidence and argument is in favour of X.*

*'**X-is-suspected**' if it is not believed but there is evidence or argument to support it.*

*'**X-is-doubted**' if (not X) is suspected or a competitor of X is suspected.*

*'**X-is-assumed**' if it is asserted by means of default values or derived by general knowledge and there is no evidence or argument to the contrary.*

These objectives are sought by Cohen **[50]** whose theory of *endorsements* is a good example of the non-numeric approach. His aim is *reasoning about uncertainty rather than reasoning with uncertainty* and to this end he develops the use of *endorsements* that are provided by information either for or against a proposition. A *ledger book* metaphor is used to represent each item of evidence either *PRO and CON*, according to the strength of endorsement. He argues that if the reasons for believing or disbelieving a proposition are provided, then one can assess whether a statement is certain enough to use as evidence. He states that *one's certainty in a result should depend on what the result is wanted for ... a proposition may be believable enough to function as a premise in an inconsequential inference, but it may be rejected for more serious purposes.*

Cohen states that *the certainty of a hypothesis can be represented as its strongest endorsement* and his metaphor of an *ordered set of increasingly stringent bureaucrats endorsing a proposal and bears a strong relation to an inductive method of reasoning.* He uses a legal analogy of proof *beyond reasonable doubt* to argue that a high probability for an event should not necessarily be a basis for action if little relevant evidence has been obtained. For example, the prior probability that a node in a

network topology requires a caesium clock may be less than 1 in 1000, but that is no basis for ruling out the option without even examining the problem. He places the *diagnostic weight* for a choice on an ordinal scale graded by the *variety of potentially relevant circumstances that fail to falsify it* that fits into systems that sequentially apply a standard set of increasingly stringent tests. He argues that a probability is based on assuming that the evidence obtained is all that is available, whereas the evaluation of a problem should be related primarily to the extent of the relevant evidence that has actually been obtained.

Fox **[62]** uses linguistic modifiers; words such as *very, relatively and approximately* to put finer meaning to otherwise similar statements. He uses these when the expert system supplies answers to the user.

There has been considerable research into on the modelling of linguistic methods on a computer system. Each technique requires the conversion of linguistic terms in to numbers for their subsequent manipulation. Unfortunately, there is considerable variation in the way different people interpret phrases and their interpretation is both contexts dependent and the degree of association the expert has with the context. For example, saying that an increase in telephone traffic is 'fairly likely' means something rather different in an economic recession than it does during periods of economic growth. Similarly, the quantitative implications of 'traffic is high' are affected by knowledge of whether the site is home for administration or manufacturing staff. A simple mapping between words and numbers are unlikely (sic) to be adequate.

A survey conducted of BT consultants on their mathematical representation of linguistic terms confirms the difficulty in utilising such in a computer system, see

figure 15. For example, whilst the average values of probable and possible were 0.71 and 0.39 respectively, the range of values had considerable overlap with the position of the words changing places on the scale of uncertainty given by some consultants.

3.5.2.2 Categoric (Logic) Modelling

Categorical modelling relies upon traditional premise and conclusions logically.

IF clock accuracy is 10^{11} THEN clock is atomic.

Whenever the conditions for a rule match the state of the inference network, the action part of that rule is exercised. The action part of this rule is:

Clock is atomic is True.

Categorical logic implementations are monotonic; that is, the addition of new information cannot invalidate any previous deductions and hence expert systems utilising a logical based system have an inability to handle conflicting knowledge, they also are only able to deal with exact knowledge.

3.5.2.3 Probabilistic Modelling

Probabilistic modelling works by virtue of weights added to data by an expert based upon his interpretation of the frequency with which the data would occur in a long series of trials. In this view the probability associated with the data is actually a property of a theoretically infinite number of sequence of trials rather than of a single event. In expert systems it is usual to distinguish between results whose

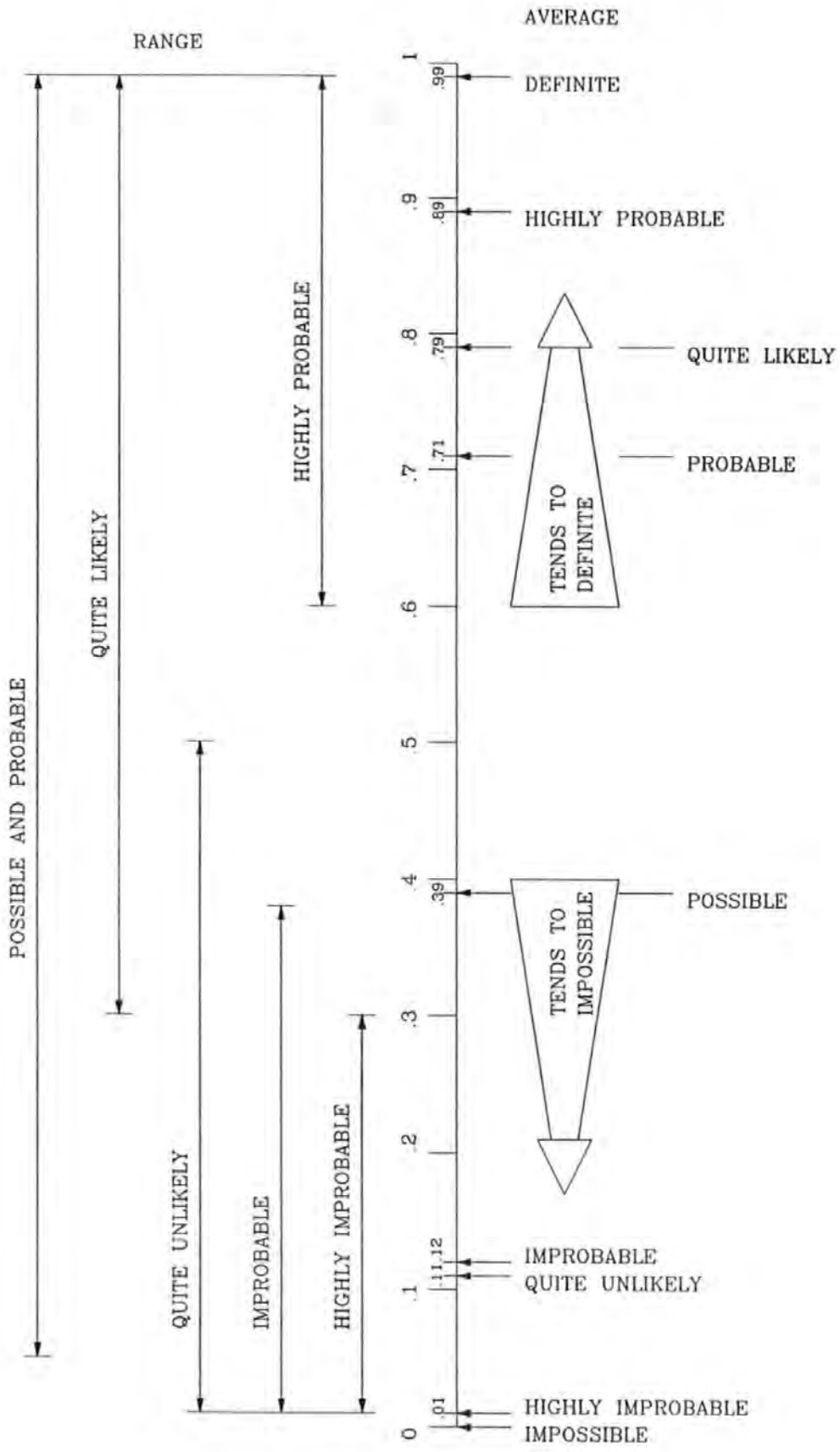


Figure 15 Survey Results of BT Consultants View of Uncertainty

probabilities are 'knowable' and have an obvious parent population and those whose probabilities are 'unknowable' and do not.

The manipulation of the probabilistic weightings are basic to axioms of probability theory, i.e.

P(a) and is the probability of 'a' being true,

P(a)[true] = 1, P(a)[false] = 0,

P(a)+P(a) = 1,

& P(a) Independent P(b).

Probabilities are not fuzzy, they are precise. They will indicate, given an infinite number of trials, the number of successful against unsuccessful attempts. Thus the probability of an outcome P(o) for the two cases AND & OR are:

P(o) = P(a) AND P(b) = P(a)*P(b),

P(o) = P(a) OR P(b) = P(a)+P(b)-P(a)*P(b).

Because probabilistic representation explicitly includes the conditioning information the representation is, as in logic, monotonic; when all probabilities are either 0 or 1 probabilistic modelling reduces to categorical modelling.

3.5.2.4 Ad-Hoc Certainty & Confidence Factor Modelling

Because one cannot be completely certain that some facts are true or that certain relations hold in the ad-hoc confidence factor modelling process, each fact and each rule can be associated with a certainty or confidence factor (CF).

The CF, traditionally, is a number in the range (-1,1). It indicates the certainty with which each fact or rule is to be believed. Positive and negative CF indicates a predominance of confirming or opposing evidence, respectively. CF of 1 or -1 indicates categorical knowledge.

Inexact reasoning is based on the construct

IF A (to degree x) THEN B (to degree y).

Such constructs are applicable to situations in which there might be more than one correct decision.

Inexact reasoning is best illustrated by the following example:

IF the site probability has multiplexers
& digital link terminates
& the site is a Transit
THEN there is suggestive evidence that the node is digital.

The numerical techniques for propagating degrees of certainty and combining evidence from different sources that were adopted in early systems such as PIP, CASNET, INTERNIST and MYCIN are described in detail by Szolovits and Pauker **[64]**. For example, in INTERNIST a disease hypothesis is seen as causing symptoms or manifestations with a certain frequency, whereas, inversely, a specific finding supports an underlying disease with some strength: as in many systems positive and negative support for a hypothesis is treated separately. Tallies of scores are used to provide control of the consultation to give focus to the crucial hypotheses **[65]**.

Other systems have attempted to provide a sounder theoretical basis for their schemes. EMYCIN (Empty-MYCIN) **[66]**, a general expert shell that grew out of MYCIN, uses the confirmation theory approach of certainty factors. For example, suppose facts a, b and c are known with CF's 1, 0.8 and 1, respectively, and two rules exist, then:

if a and b then d (CF = 0.6);
if c then d (CF = 0.8).

This is represented in figure 16.

The certainties of a and b are combined by the 'fuzzy' operator $CF(a\&b) = \min[CF(a), CF(b)] = 0.8$. In turn, this attenuates the certainty propagated to d into 0.8 x 0.6 = 0.48. With both rules supporting d, the certainties combine with the function $1\text{-}CF(d) = [1 - CF(c)]^* [1 - CF(a \& b)] = 0.896$, thus reinforcing each other.

PROSPECTOR, by contrast, operates with variables that represent propositions which may be true, 'x', or false, 'not-x'. Rules are expressed as

if X then Y (to degree LS,LN),

where LS is the likelihood ratio p(x | y)/p(not-x | y), which measures the multiplicative increase in the odds on Y being true if X is true, and LN = p(not-x | not-y)/p(not-x | y) is the effect if X is false. However, although specification of p(x),p(y), LS, and LN should obey the laws of probability, these four assessments are not necessarily forced to

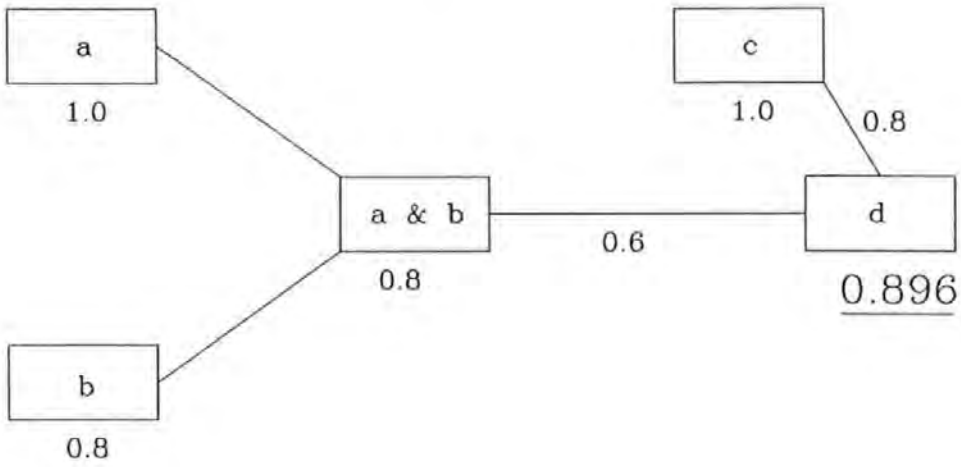


Figure 16 Combination of EMYCIN Rules

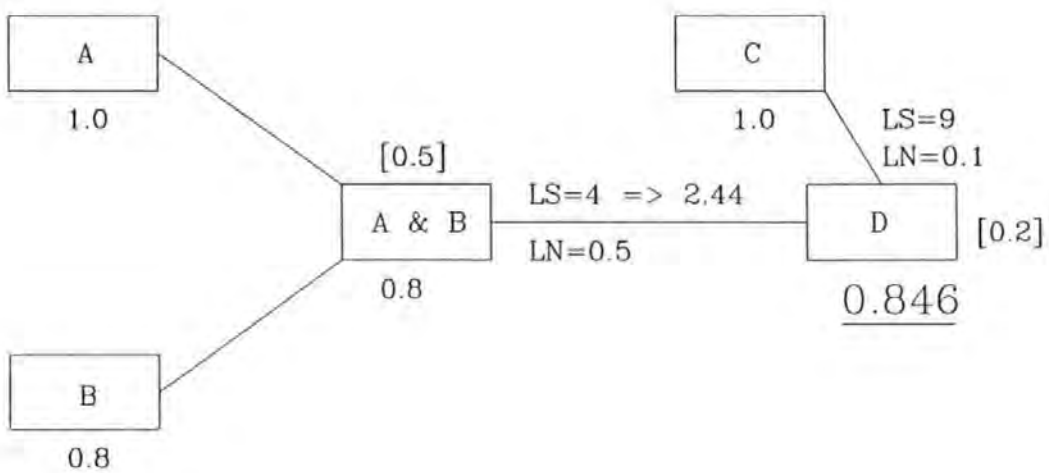


Figure 17 Combination of Prospector Rules

cohere. Instead, the calculus is adjusted to allow for non-coherent probability assessments. For example, figure 17 represents the rules:

if A & B then D (to degree 4, 0.5);

if C then D (to degree 9, 0.1);

together with the prior probabilities displayed in square brackets where external information has changed the probabilities of a, b and c to 1, 0.8 and 1 respectively.

The probability of the conjunction a & b is determined by the fuzzy minimum of its components, which may be considered as an upper bound on the possible values p(a&b) could take on. Just as in MYCIN, having some uncertainty about A & B attenuates the impact, LS, of the rules. This is done by calculating an effective likelihood ratio in the following manner. If A & B were at its prior probability p(a&b) = 0.5, then p(d) would still be set at 0.2. If A & B were known to be true, then p(d | a&b) would be 0.5 (since the prior odds of 0.2/0.8 would be multiplied by LS = 4 to give odds of 1).

PROSPECTOR interpolates between these fixed points to say that if evidence E makes p(a&b | E) = 0.8, then the posterior probability of D = d should be the fraction [p(a&b | E)-p(a&b)]/[1-p(a&b)] = 0.6 of the way between 0.2 and 0.5, or 0.38. This corresponds to posterior odds of 0.61 or a likelihood ratio of 0.61*4= 2.44. Thus LS = 4 has been attenuated to 2.44 by the uncertainty concerning A & B. To combine the effect of the two rules, there is an assumption of independence of C and A & B conditional on D, so that the effective likelihood ratios are multiplied to give final odds on D = d of 9*2.44*(0.2/0.8) = 5.49, or a probability of 0.846.

The examples illustrate two points.

Firstly, in attenuating the strength of a rule due to uncertainty concerning the premise, neither system takes into account the implication of the premise being false, even though that information may be available. Brooks et alii **[67]** point out the problems this brought in implementing EMYCIN. Suppose two rules exist:

if Fault 1 then Remedy 1 (CF = 0.6);
if Fault 2 then Remedy 2 (CF = 0.9).

Suppose CF (Fault 1) = 0.9 and CF (Fault 2) = 0.7, then attenuating the certainty factors by multiplication reveals Remedy 2 as the preferable action although Fault 1 appears most probable. The problem is that no indication is given of the support for Remedy 1 if Fault 1 is not true.

Secondly, the procedures in EMYCIN and PROSPECTOR seem to have little to do with each other when combining the impact of rules. However, Hajek and Havranek **[68]** has defined a unified treatment of combining functions. Specifically, he shows that *any reasonable combining function on 'certainties' measured on a scale (-1,1) ... to the additive group of real numbers on (-inf,inf),* which he calls *a weight of evidence scale*; that is, two 'certainties' can each be transformed to a number, the numbers added, and the result transformed back to the resulting combined certainty.

In ad-hoc confidence systems, e.g. MYCIN/EMYCIN, uncertainty is viewed as a degree of confirmation. The degrees of confirmation are numeric values in the interval [-1,1] where 1 means that the evidence confirms the hypothesis, -1 means that the evidence un-confirms the hypothesis, 0 means that the confirming evidence balances the un-

confirming evidence, and an intermediate value means that the evidence is partially confirmed (or un-confirming) by the evidence.

The advantages of ad-hoc system are that the calculus has linear information complexity and linear time complexity. The disadvantages include the fact that it is possible to get a different CF associated with conclusions dependent upon the order in which the rules are invoked! There is evidence that the assignment of CFs to rules is based upon the seriousness of the consequences. Thus the true significance of CFs was more akin to measures of importance rather than uncertainty, i.e. confusing usefulness and uncertainty. In MYCIN CFs are treated mistakenly as similar to, or identical with, the belief or disadvantage attached to each set of alternative courses of action.

3.5.2.5 Bayesian Modelling

Although most statisticians would support the use of probabilistic reasoning to model the chance effects of underlying hypotheses, many would speak against the use of probability to express uncertainty about events that could not be regarded as either having been generated by some random process or the development of information as time progresses. They argue that there is no such thing as 'the' probability of an event: different people or the same person at different times may legitimately assign different probabilities to the same event; that is, it is a 'personal' probability rather than 'the' probability. This subjective view of probability does not mean that probability assignment is arbitrary because if they are legitimate they must be consistent with the axioms of probability; given an infinite number of attempts then the subjective probability would revert to frequency probabilities.

In essence, Bayes theorem **[38]** states:

Posterial probability = k*(Prior probability)*(Likelihood)

where likelihood is the probability of the evidence given the hypothesis.

Bayesian inference is essentially recursive where the posterial probability becomes the new prior probability of the hypothesis when the theorem is revised to compute a new posterial probability given new evidence. If the new evidence is irrelevant the posterial probability equals the prior probability.

In Bayesian calculus, uncertainty is viewed as a probability that can be interpreted as a relative frequency, as a degree of belief or in some other manner. The probabilities are numerical values in the interval (0,1) where 1 means that an event always occurs, 0 means that an event never occurs, and an intermediate value means that an event sometimes occurs.

The advantage of the Bayesian approach is that the computation has well understood mathematical properties.

However, the disadvantages include:

an exponential number of a-priori probabilities are required which inevitably leads to the imposition of simplifying assumptions that may or may not be valid in a given domain;

statistical analyses required to determine the prior probabilities tend to require a massive amount of data that may not be available;

a person is often not perfectly calibrated, i.e. when for all events or propositions to which they assign a given personal probability, P, the proportion that occurs correct is not always P.

3.5.2.6 Fuzzy Logic Modelling

Expert systems often employ a brand of logic, developed by Zadeh, called fuzzy logic **[69 & 70]** in which truth values instead of being limited to either 1 or 0 are allowed to be qualitative, i.e. they can take the form true, not true, very true, more or less true, false, very false, etc. and the rules of inference are approximate rather than exact.

The premises are assumed to have the form of fuzzy propositions, e.g. 'X is much smaller than Y is quite true' or 'if X is small is possible then, Y, is very large is very likely', etc. By these means, a possibly imprecise conclusion can be deduced from a collection of imprecise premises. Such reasoning is for the most part qualitative rather than quantitative in nature, and almost all of it falls outside of the domain of applicability of classical logic.

Zadeh has claimed that *the imprecise language that characterises much expert knowledge argues for the use of fuzzy reasoning, in which one can quantify the extent to which a proposition is true.* Adlassnig **[71]** and Adlassnig et alii **[72]** and Fieschi et alii **[73]** describe applications of this approach, which aims to allow linguistic terms such as 'high' to be directly entered into an expert system without further definition.

Gaines **[74]** provides a discussion of the distinction between fuzziness and probability. As expert systems have to work with uncertain knowledge in the sense that empirical regularities are usually only valid to a certain degree, IF-THEN rules have to be

modified either by adding a probabilistic qualification to their logical interpretation in the form:

Rule No: IF A THEN C & D WITH CERTAINTY F;

or by defining new logical relationships such as 'pessimistic AND' & 'Optimistic OR': the de-facto labels being CAP & CUP respectively.

Inference now tasks the form:

Rule No: IF A & B THEN B & C, CUP, B & D.

Fuzzy-set modelling and manipulation utilise the basic axioms of traditional set theory as detailed below:

Complementation

In classical set theory, the complement of a set contains all things that are not in the set. In a normal set with a sharp, crisp boundary, its complement is everything outside the boundary. Similarly, the complement of a fuzzy set is defined as that set whose grade of membership is exactly one minus the grade of membership of the original set. For example, consider the representations of 'tall' and 'not tall' in figure 18.

Being a member of 'not tall' depends only on 'tall'. The usual complementation is recovered when fuzzy sets are crisp, that is when all membership degrees are equal to 1. Also, membership in 'not tall' becomes smaller as membership in 'tall' increases.

Tall		<----->	Not Tall	
5'0"	0.00		5'0"	1.00
5'4"	0.08		5'4"	0.92
5'8"	0.32		5'8"	0.68
6'0"	0.50		6'0"	0.50
6'4"	0.82		6'4"	0.18
6'8"	0.98		6'8"	0.02
7'0"	1.00		7'0"	0.00

Figure 18 Complementation

Tall AND Short	
5'0"	0.00
5'4"	0.08
5'8"	0.32
6'0"	0.50
6'4"	0.18
6'8"	0.02
7'0"	0.00

Figure 19 Intersection

Not Tall		Not Short		Middle-sized	
5'0"	1.00	5'0"	0.00	5'0"	0.00
5'4"	0.92	5'4"	0.08	5'4"	0.08
5'8"	0.68	5'8"	0.32	5'8"	0.32
6'0"	0.50	6'0"	0.50	6'0"	0.50
6'4"	0.18	6'4"	0.82	6'4"	0.18
6'8"	0.02	6'8"	0.98	6'8"	0.02
7'0"	0.00	7'0"	1.00	7'0"	0.00

Figure 20 New Information

The above does not determine complementation uniquely, even if a certain change in the membership value of 'tall' has the same effect on the membership of 'not tall'. Justification becomes even more difficult by trying to generalise the definition of a fuzzy set.

However, by the complementation process, it is possible to generate another fuzzy subset, 'not tall', from the original fuzzy subset 'tall'. Now a new label can be assigned to the new fuzzy subset, for example, the label 'short'.

Intersection

The intersection of two crisp sets contains those elements of the sets that are both in one and in the other. But, by definition, an element of a fuzzy subset may be partly in one set and partly in another. Therefore, it is as true that an element is in the intersection of two fuzzy subsets as it is true that it is in either one.

Thus, at any point in the domain, the truth of the proposition that an element is in the fuzzy subset 'tall AND not tall' is the minimum of the truth of the propositions that it is in 'tall' and that it is in 'not tall', i.e. tall AND not tall = min (tall, not tall).

Figure 19 illustrates the Intersection of the two fuzzy subsets 'tall' and 'short'.

What the above says is that in this example there is a person that is 'tall and short' which means they are middle-sized. Thus, it is expected that the highest membership degree to be in the middle of the domain, and the lowest to be at the edges.

In practice, it is difficult to imagine using the phrase 'tall or short' to describe a persons' height. The description 'not tall' and 'not short' seems more reasonable. From the fuzzy set point of view the representations are equivalent.

More interestingly, one can derive new information by intersection. For instance, from 'not tall' and 'not short' means 'middle-sized', see Figure 20.

Union

From the above, it is now possible to define the concept of Union, i.e. the set of those elements that belong to either one or both of the constituent sets. In the domain of two fuzzy-sets, the membership of the Union of the fuzzy subsets cannot be less than the membership of either component, i.e. tall OR short = max (tall, short)

Figure 21 illustrates Union applied to the fuzzy subset 'tall' and 'short'.

The membership function attains its highest values at the edges of the domain, and its smallest value at the centre. This illustrates what is meant by saying that a man is 'tall or short', means he is 'not middle-sized'.

It is now possible to redefine this new fuzzy set, for example, from 'tall' or 'short' infers 'not middle-sized', see Figure 22.

In summary, the rule of fuzzy set manipulation follows those of crisp set manipulation. and uncertainty is viewed as a degree of set membership. The degrees of membership are numerical value in the interval (0,1) where 1 means that an object is a member, 0 means that an object is not a member, and an intermediate value means that an object

Tall OR Short
5'0" 1.00
5'4" 0.92
5'8" 0.68
6'0" 0.50
6'4" 0.82
6'8" 0.98
7'0" 1.00

Figure 21 Union

Tall	Not Short	Middle-sized
5'0" 0.00	5'0" 1.00	5'0" 1.00
5'4" 0.08	5'4" 0.92	5'4" 0.92
5'8" 0.32	5'8" 0.68	5'8" 0.68
6'0" 0.50	6'0" 0.50	6'0" 0.50
6'4" 0.82	6'4" 0.18	6'4" 0.82
6'8" 0.98	6'8" 0.02	6'8" 0.98
7'0" 1.00	7'0" 0.00	7'0" 1.00

Figure 22 New Information

is a partial member. The advantages are low information and time complexity and applicable to lexical translation. The disadvantage includes construction of set memberships which are not clear and its operator definitions lack mathematical rigour. The literature gives little guidance on usage.

3.5.3 Limitations of Existing Models

Uncertainty is the trigger that motivates reason.

Thus when the brain, whilst assimilating information received finds an exception to a rule, for example a fact that cannot be proved, or disapproved, or data compiled or collated by persons long deceased, whereby no questions or queries can be asked, uncertainty arises and induces a line of reasoning to discover a true and certain solution. This line of reasoning is logical and is largely based upon the brain's experience, whereby another dimension is incurred that of intuition.

To repeat artificially this sequence of events, a multi-dimensional method of dealing with uncertainty must be built into any mechanical method to be used.

Current techniques for dealing with uncertainty are ad-hoc, crude and singular in dimension. These methods may be sufficient for a rank ordering, the current practice, because order statistics are more robust than interval statistics. Techniques are needed that can weight data along with costs, benefits, and risks. It doesn't matter if a resultant telecommunication network design has a 75% chance of using analogue technology if there is also a 10% chance that there is a need to support digital link or be an (ISDN), so a rank order of the likelihood of hypotheses is not useful.

Concern also centres on the propagation of uncertainty, with opposing views of symbolic manipulation over numerical representation. The former proposes to provide a formal logical basis rather than a formal numerical system requiring formal definitions of terms used to talk about uncertain events. It should be pointed out that even if the statistician's viewpoint that probability is probably (sic) the best way to represent uncertainty is correct, it is still not necessarily the best means for communication with most people. The representation of uncertainty has reflected a movement away from logic, judgement, and opinion in favour of calculation.

Overall, current methods of uncertainty pose problems in the following ways.

3.5.3.1 Lack of Quantitative Input Data

Even if a probabilistic representation is theoretically reasonable, it may be that there are either insufficient past data or insufficiently confident experts to assess anything other than an approximate or imprecise figure, particularly concerning relationships between more than two variables.

3.5.3.2 Ignorance or Uncertainty?

Cohen **[50]** states that *the Bayesian view of probability does not allow one to distinguish uncertainty from ignorance*, for example, if a rule suggests that either answer '1' or answer '2' is best, but does not provide any basis for choosing between them, a probability approach will assert the prior probability while it is often argued that a range of probability may be appropriate.

3.5.3.3 Perception Colours Judgement

Kahneman et alii [75] reveal the way perception colours our judgement. The slightest difference in the way evidence is presented to us, what we see first, last, etc., can alter the way we perceive a situation. They identify many factors that affect our judgement the most significant being 'Anchoring'. For example, given a task to multiply a set of numbers in five seconds either 8*7*6*5*4*3*2*1 or 1*2*3*4*5*6*7*8, whilst the two strings are clearly equivalent, the estimated answer to the first string is generally higher than the second! Anchoring means the uncertainty we attribute to a fact, etc. depends upon where it is placed in a string of other facts.

3.5.3.4 Explanation & Assessment of Conclusions

It has been said that for probabilistic systems, *there is an unavoidable loss of comprehensibility to the person using them ... when the list of rules is long, it may not be clear how each of them (or some combination of them) contributed to the conclusion,* and that *useful explanation for the user must involve symbolic reasoning and the provision of a logical argument ... also all numeric calculus aggregates numbers and keeps no record of divergence in opinion,* [61]. Unless these have some interpretation, it is difficult to see how a group of developers can agree on reasonable numbers, for a discussion on the assessment of low certainty factors see Shortliffe and Buchaman [56].

3.5.3.5 Qualitative & Quantitative Learning

Hink and Woods [76] identify three ways experts view uncertainty: *perception, probabilistic judgement and choice.* The uncertainty calculus does not encompass and manipulate these diverse attributes. There is also the need to be able to update the numbers in the light of experience.

3.5.3.6 Uncertainty & Knowledge

Uncertainty terms are used uniformly for the different kinds of knowledge - facts, rules and data. It is important to distinguish between the different types and sources of uncertainty since they need to be treated in different ways.

3.5.3.7 Single Dimensional Uncertainty

Uncertainty as current defined is singular in dimension. It does not take account of the fact that the different types of uncertainty need to be promulgated through the system in different ways. Indeed, some attributes of uncertainty may be regarded as mutually inclusive whilst others mutually exclusive.

3.5.3.8 Inconsistency

Magill et alii [77], Castillo et alii [78] and Garbolino [79] have shown that the uncertainty techniques implemented do not follow strictly the theories upon which they were based.

3.5.3.9 Quantifying 'Unknown' - '0' or '1' ?

It is usual to allocate either a zero or one to the uncertainty value dependant upon the uncertainty calculus used. Traditional logic demands a unity value for unknowns, any zero value used in the calculus of uncertainty will result in a resultant of zero. Confidence Factor calculus requires zero values for unknowns as maximum values are dominant in the calculus.

3.5.3.10 Inference Under Uncertainty

Inference under uncertainty with an expert system has certain crucial distinctions from the general problem of scientific inference.

Firstly, the system is bounded: there is only a limited set of facts concerning the current problem that are available to it and all its reasoning must be based on those facts and those alone.

Secondly, the system knows exactly what questions have not yet been asked, and so at any time it is precisely aware of the limited basis on which it needs to form a judgement.

3.6 Uncertainty and Expert Systems

3.6.1 Information Takes the Form of Rules, Facts & Data

The ability to make a decision with uncertain knowledge is central to expert systems. They make decisions between alternative options. Decisions are made in adopting or eliminating hypotheses in selecting strategies for problem solving. Every heuristic,

and every inference in an expert system, is ultimately concerned with a decision to follow some path in preference to others based upon the analysis of weightings.

On this view, decision making in an uncertain environment is as pivotal as, say, logical inference or knowledge representation. It is curious, therefore, that it has not been regarded as a central problem. When uncertainty procedures are used, they are often ad-hoc or an inheritance from statistical decision theory.

Nothing is more certain than uncertainty, a paradox in terms, but consider: the human mind is presented with a multitude of information from which, at time to time, must arrive at a result or conclusion. This information is usually in the form of three definitives, i.e. a rule, or rules, facts and data. Data can be uncertain, facts can be unreliable, and rules can be rough and ready.

Uncertainty implies doubt. Even a-priori statements, if they are true and thus unassailable, are not necessarily immune from doubt for it is possible to make a mistake and believe an a-priori statement to be true when it is not.

This implies that data, which are statements about a particular situation, but also facts and rules, which are universal knowledge statements, are all susceptible to doubt and hence uncertainty.

To discuss and define uncertainty one must begin at the other end of the spectrum and look at certainty. Absolute certainty can only be measured in a mechanical fashion, and this in itself is fully dependent upon 100% efficiency and control of the human senses required for the use of any such. These senses being, sight, hearing,

touch and voice. Such clarity is not possible when the problem to be solved can take on an infinite number of inputs and outputs.

3.6.1.1 Rules

All things living, dead and inanimate are governed by rules, unfortunately, or perhaps fortunately, there are exemptions to these rules, where things happen in an illogical and haphazard way; this is where uncertainty arises, the brain must then draw upon experience to determine a solution or procedure.

In the MYCIN system probabilities, or 'certainty factors' were represented as numbers attached to rules. In the following example, the conclusions BT and Mercury are quantified with values 0.8 and 0.2:

IF 1) the site of the node is a Primary Centre
2) the node is a transit, and
3) the node is terminated in digital links

THEN there is evidence that the terminating links are BT(0.8) or Mercury(0.2).

In this case, the certainty factors are intended as probabilities, and this is evident here. Unfortunately, in the literature no guidance is given as to the meaning of a certainty factor. Is it a probability for the conclusion as was intended in MYCIN? Is it an assessment of confidence in the inference? Or does it represent the importance of the conclusion? These are all possible uses because the logical process is treated as independent of the 'statistical' process and hence cannot place any restrictions on the interpretation of the numbers.

(The situation is further confused because certainty factors play a part in MYCINs control of inference, i.e. they have a non-logical, non-statistical role.)

3.6.1.2 Facts

If a fact is there, or was there, the mind can qualify the certainty of the fact by using all senses necessary thus leaving no room for uncertainty: having to make decisions on facts that were determined by another and having to rely upon these without proof becomes uncertain.

The CASNET system dealt with uncertainty by calculating weights of evidence using parameters associated with the facts in its knowledge-base. For example, a causal fact is represented as:

> Primary_Master_Site_Location causes
> Installation_Of_Digital_Node(0.7)

Given this fact, the observation of a Primary Master is evidence for digital node, and vice versa. Like MYCIN, CASNET incorporated a technique for revising its belief in hypotheses, e.g. belief in the hypothesis 'digital node' rises and falls as data are accumulated. The empty shell that was developed from it, EXPERT, also used this technique. However, the meaning of these numerical coefficients is quite unclear. Are they probabilities? Are they measures of strength of the causal association?

3.6.1.3 Data

Data are a collation of figures and opinions. If the figures are of recent origin and can be verified, then uncertainty would only arise in considering the sense used to obtain the data, how old, how efficient, etc. Opinions, the strength of which would depend upon the individual giving them an experienced brain, would accumulate.

Data, of long past origin where no verification is possible, gives rise to uncertainty.

The term data is used here to refer to data gathered while solving a problem or making a decision. Data, which are statements about particular situations, are distinct from knowledge like facts and rules which might be said to be universal statements and always true:

EMYCIN: Has the site got a digital node?

⇒ Yes (0.3).

The number is treated as a certainty factor and enters the inference engine in the normal way. But again, whatever the designers meant by a certainty factor can one be sure that the user is expressing certainty in the sense? The user may only wish to record a judgement on the reliability or credibility of an individual statement, say, without reference to the frequency of such statements. There is no way of ensuring that the sense in which the certainty factor is used does not differ from the sense in which it used in other situations or on different occasions.

3.6.2 Graphical Communication of Uncertainty

With the exception of work by Ibrekk and Morgan **[80]** there have been few evaluations of the effectiveness of graphical means of communicating uncertainty. They use a number of statistical distributions including error bars that span a 95% confidence interval, probability density functions, charts and cumulative distribution functions. Their conclusions include: *the study clearly needs more study using a variety of improved and more focused designs ... that rusty or limited statistical knowledge does not significantly improve the performance of semi-technical or lay people in interpreting displays that communicate uncertainty.*

Larkin et alii **[81]** show that a figure can be superior to a verbal and numeric description of solving problems. They argue this is because;

figures can group together all information that is used together, thus avoiding large amounts of search for the elements needed to make a problem-solving inference;
figures typically use location to group information about a single element, avoiding the need to match symbolic labels;

figures automatically support a large number of perceptional inferences, which are extremely easy for humans;

to be useful a figure must be constructed to take advantage of these features and ... users know the appropriate computational processes for taking advantage of them.

Thus, the suitability of the display depends upon the training and experience of the user. However, it is probably unwise to design displays for the least skilled, as

exposure to sophisticated displays will help the user to improve the skills in interpretation and analysis.

The key advantage of diagrammatic over numeric representation of data is its computational edge. That is a figure can be a better representation not because it contains more information but because indexing of this information can support extremely useful and efficient computational processes.

When a person looks at a graph, the information is decoded by their visual system. A graphical method is successful only if the decoding is effective. Cleveland et alii **[95]** have ordered graphical methods by perceptional value as shown below:

position along a common scale;

length;

angle;

slope;

area;

volume;

colour.

The order provides a guide to displaying data so that the result allows for more effective perceptional analysis.

CHAPTER 4

BUILDING AN EXPERT SYSTEM

4.1 Introduction

This chapter discusses how expert systems are developed. Particular emphasis is placed upon their use in telecommunication network design and the impact of uncertainty upon that design.

McDermott [83] has suggested the following four-stage strategy for introducing expert system technology into industry:

isolate a set of real problems in an industrial environment;

select those problems whose solutions appear to require a large amount of knowledge or involve a great deal of search, (the other problems will be of the number-crunching variety which can be dealt with by existing software tools);

build simple systems which begin to solve these problems and which can be used in an industrial environment;

adapt and extend these systems to make them more intelligent.

4.2 Real Problem Selection

Solving the problem of telecommunication network design, when approached holistically, can be seen to be a potentially infinite task. Figure 23 outlines just a few of the interdependent factors that need to be considered before completing such a design and shows the complexity of the task. The expert system must, therefore, break down the problem into manageable tasks, each with obtainable goals and each goal in turn contributing to the solution of the problem. Figure 24 represents the problem as a peripheral area that itself is composed of smaller area sub-problems. This approach gives each module a single interface to the next, making each capable of being tested and verified. Typically within the goal of network design, sub-problem could comprise Network Signalling, Synchronisation, Routeing, Numbering, etc.

The expert system developer focuses on the experts' experiential knowledge, including their assumptions, definitions, strategies, rules-of-thumb and intuitive additions. To facilitate this task, the developer usually presents the experts with a selection of hypothetical problems in which the relevant factors are systematically varied. The experts are encouraged to describe the rules they use to arrive at their conclusions. They are then asked to try and analyse these rules at a more detailed level to uncover the data elements from which they are constructed. The choice of what level of detail is needed to describe these basic data elements may initially be just intuitive guesswork on the part of the interviewer. Later, as the concepts and rules needed to answer the hypothetical questions are refined, the interviewer and expert together may discover the level of representation best suited to the application in hand. Defining the basic concepts in the domain and representing them in an explicit and precise fashion is a difficult process and one not well understood at present. At the

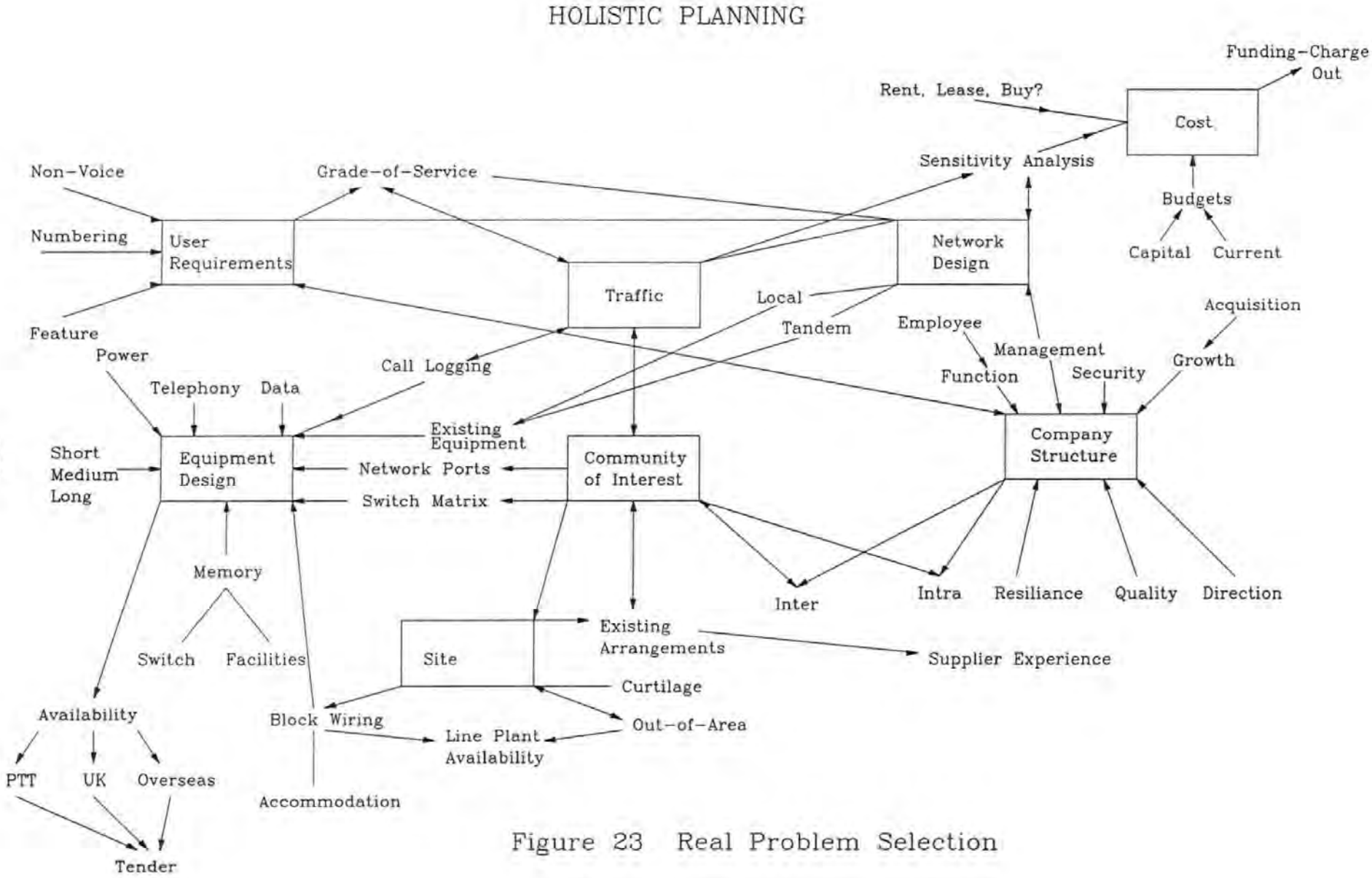


Figure 23 Real Problem Selection

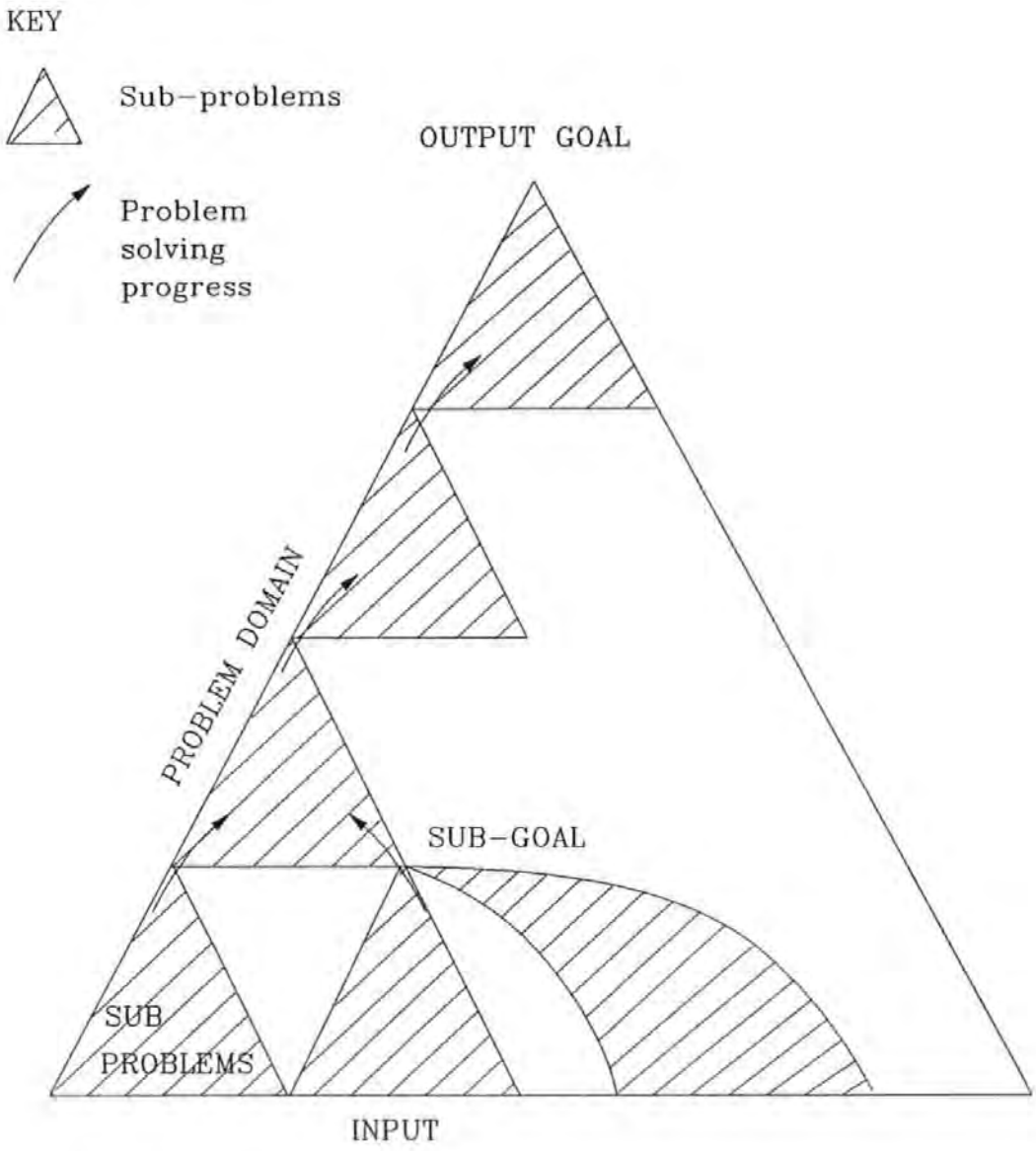


Figure 24 Problem Partitioning

same time, the decisions made at this stage are critical because they greatly influence the potential capabilities of the system, including its flexibility and powerful.

Alongside the interviews, other knowledge acquisition methods are usually employed, for example:

studying past and present problems and their solutions;

identifying idealised problem types and associated solutions;

carrying out literature searches;

systematic observation of experts at work in a real environment, solving actual, as opposed to hypothetical, problems; (this is vital in order to distinguish between the knowledge that experts think they use and the knowledge they actually use in solving problem;

self-analysis (the expert consultant).

Once all this knowledge has been extracted and formalised, a prototype expert system can be implemented. This system must then be rigorously tested to see if it is capable of predicting the expert's decisions on both old and new design problems. Further interaction with the expert will be necessary at this stage to refine the system to make it model the expert more closely.

4.3 Implementing Knowledge as a Rule-Based System

Having acquired the knowledge necessary to build an expert system in a particular domain, two important design decisions follow **[84]**:

how is the knowledge to be encoded so that it is both usable by the system and readily comprehensible to the human user?

how can this knowledge then be used to solve problems in the domain efficiently? (i.e. what should the control structure of the system be?).

4.4 Encoding Knowledge

An expert system, as already identified, comprises facts, data and inference rules: there are many different ways of representing this knowledge, the most common is to encode the knowledge in the form of a 'production system'. A simple example of a 'production rule' would be:

IF cond1, cond2 ... condN THEN action1, action2, ... actionN in which cond1, etc. are assertions about the knowledge under discussion.

In a full-scale production system, a whole set of these rules are chained together to describe complex processes. The condition part of each rule refers to some fact in the database that will be either true or not at any time. When all its conditions of a rule are true then the action part of that rule may either specify changes to be made to the database, e.g. inferring new data, or may specify which rule should be considered next

or may specify some action to be performed in the outside world. *A production system thus conceptually goes through an evolutionary process. At each cycle, one rule whose conditions are satisfied is selected and activates, its actions will then affect which rules can be activated on the next cycle* **[84]**. In this way a complex decision process can be described.

The number of rules in most working expert systems is around a hundred and increases slowly in relation to the complexity of the problem domain. One difficulty a designer will always face is knowing when the rule set is complete.

Production rules are not the only way to represent knowledge: others are as follows.

Frames. A frame is a description of an object that has different 'slots' available, each containing a particular piece of information about that object. In addition, slots can contain procedural information, or default values that add to the power of this formalism. An example of a frame is:

FRAME NAME:	Switching;
TYPE:	DMS 100;
MAKE:	British Telecom;
SIZE:	1000 ports.

Object-Attribute-Value-Triplets. Attributes, or general characteristics, are given values and assigned to a particular object, for example NT-Switch-Digital. This representation of knowledge is popular for describing objects in the rule base, i.e. representing facts about a domain.

Semantic Networks. These are a more general form of the graph or tree, whereby objects are 'linked' together using nodes and arcs according to their relationships to each other or to various features that they have in common. The telecommunication network is well represented in this form.

Logic. The most usual logic systems are prepositional logic and predicate calculus, for example:

Network Centre(x) + Digital(x) $\Rightarrow$ Primary Master(PM) Site(x),
i.e. If x is network centre node and x is digital then x is good PM site.

Because it is possible to encode the knowledge in a problem domain in so many different ways, it is vital that designers take into account the following attributes in any expert system design.

Consistency - semantically equivalent concepts should have similar representations.

Simplicity - the highest-level (coarsest grain) representation possible should be chosen to reduce the number of basic concepts and thus keep the number and size of the rules to a minimum. This is important for clarity and efficiency.

Modularity - rules should capture discrete pieces of knowledge so that they may be added or removed easily without destroying the logic of the system.

Flexibility - so that rules can be added without the whole structure growing unwieldy; important for user intelligibility.

4.5 Implementation Languages

In choosing the implementation language for an expert system, the designer has two main choices, either to use a conventional, algorithmic, high level language, e.g. Pascal, C, etc. or to use an AI specialist language, e.g. Lisp or Prolog. Both of the latter have distinct advantages over conventional languages when it comes to representing and manipulating human knowledge; for review, see Charniak **[85]**. In the past, Lisp has been the most commonly used language of the two for AI applications but now Prolog is rapidly growing in popularity. This is probably because many of the features needed in expert systems are incorporated within the language. Prolog is also the language that the Japanese have selected for implementing their fifth generation computer system.

Prolog is a rule-based programming language based on predicate logic (Kowalski **[86]**, Clocksin and Mellish **[87]**). The most important difference between Prolog and other conventional languages is that the program is written not as a set of instructions to be obeyed but rather as a set of statements of symbolic logic. There are two fundamental types of statement - facts and rules.

An example of a Fact is as follows :

x in-class y

where 'in-class' is a relation name. This fact tells us that the class 'x' are in the class 'y'.

An example of a Rule is as follows:

x in-class z if x in-class y and y in-class z

This reads 'a class x is in the class z if it is in some class y that is in-class z'. This tells us that 'in-class' is a transitive relation.

Prolog computation is a deduction from the facts using the rules, i.e. the facts are the database that is interrogated by the rules. This declarative nature of Prolog makes it ideal for expert system applications, many of which can be viewed as deductive use of data. However, since few applications fall solely into the domain of Artificial Intelligence (AI), expert systems should allow the Prolog to access into conventional languages for such application as arithmetical processing, etc.

4.6 Using Knowledge to Solve Problems

Having encoded knowledge about this problem domain in the form of facts and rules, how can this knowledge be manipulated to solve problems in that domain? Two different control strategies which an inference engine may use are based on two fundamentally different ways in which humans reason.

Data-directed processing: sometimes humans reason in a data driven, event driven, bottom-up or forward direction; i.e. they start from the available information and try to draw conclusions that are appropriate to their goals. For example, in a fault diagnostic task, an engineer might have no idea what is wrong with a piece of equipment, so he may work up through the set of measured evidence until he has narrowed down the set of possible faults that can be inferred.

Goal-directed processing: at other times humans reason the opposite way in a goal driven, expectation driven, top-down or backward direction, i.e. they start with a desired goal or expectation of what is to happen and then they work backwards through the available data looking for evidence that supports or contradicts their expectation. For example, an engineer may expect from the start that the equipment has a particular fault and he therefore works backwards from this hypothesis to discover which evidence needs to be asserted for the hypothesis to be supported.

Which of these two inference methods is employed in present expert systems depends upon the problem domain involved. In the case of human reasoning, however, data-directed processing is normally flexibly mixed with goal-directed processing. This means, for example, that an engineer solving a fault problem may be using goal-directed processing because he has a hunch that the equipment has a particular fault, he may switch automatically to data-directed processing if an unexpected, but relevant, piece of information or evidence is suddenly found. This is the sort of flexible control structure that designers of expert systems seek to emulate.

At the systems level this data and goal directed processing involve a basic control cycle called, say, identification or action cycle. In the case of a data process, the identifier part of the cycle consists of looking for matches between the data in the production rules and the database (e.g. sub symptom U and sub symptom V in Rule X - see figure 25). The action part then consists of selecting one of these rules and using it, thereby changing the database and so affecting the rules that could possibly be used on the next cycle.

In the case of goal processing the identifying part of the cycle consists of looking at the action side of the rules in the database (e.g. fault Y and fault Z in Rule Y) to find ones that match the current goal. The 'act' part of the cycle then consists of selecting one of these rules and looking at what conditions would make if fire and then looking back for other rules (e.g. Rule X) whose action parts conclude these conditions and so on in a recursive fashion. The expert system should also be able to discard much of the knowledge it generates as soon as it is determined to be useless to the user.

Conditions	Actions
RULE X **IF** sub-symptom U, sub-symptom U	**THEN** symptom W
RULE Y **IF** symptom W, symptom X	**THEN** fault Y, fault Z.

Figure 25 Sample of Two Possible Rules in a Diagnostic System

The procedural language method of identifying the end-goal has to be broken down into manageable units by looking for 'natural' divisions between:

the different tasks to be performed;

the knowledge being used.

Figure 26 indicates a free engine structure that allows for full flexibility, ease of programming and the ability to develop each sub-goal in turn. The contents of each sub-goal can now view requisite knowledge in the form of a spider diagram: figure 27 illustrates this point with a simple example: The Maintenance Help-Desk.

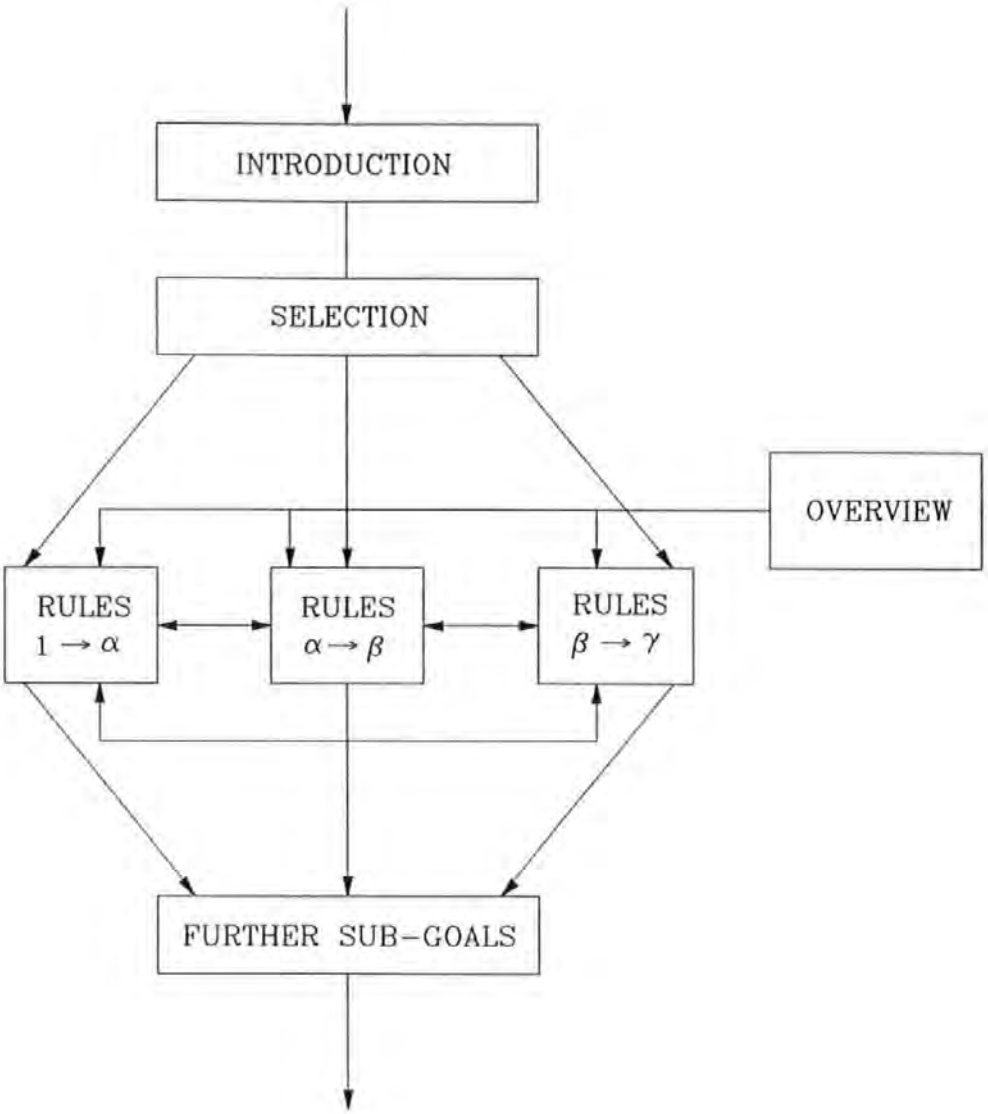


Figure 26 Free Engine Structure

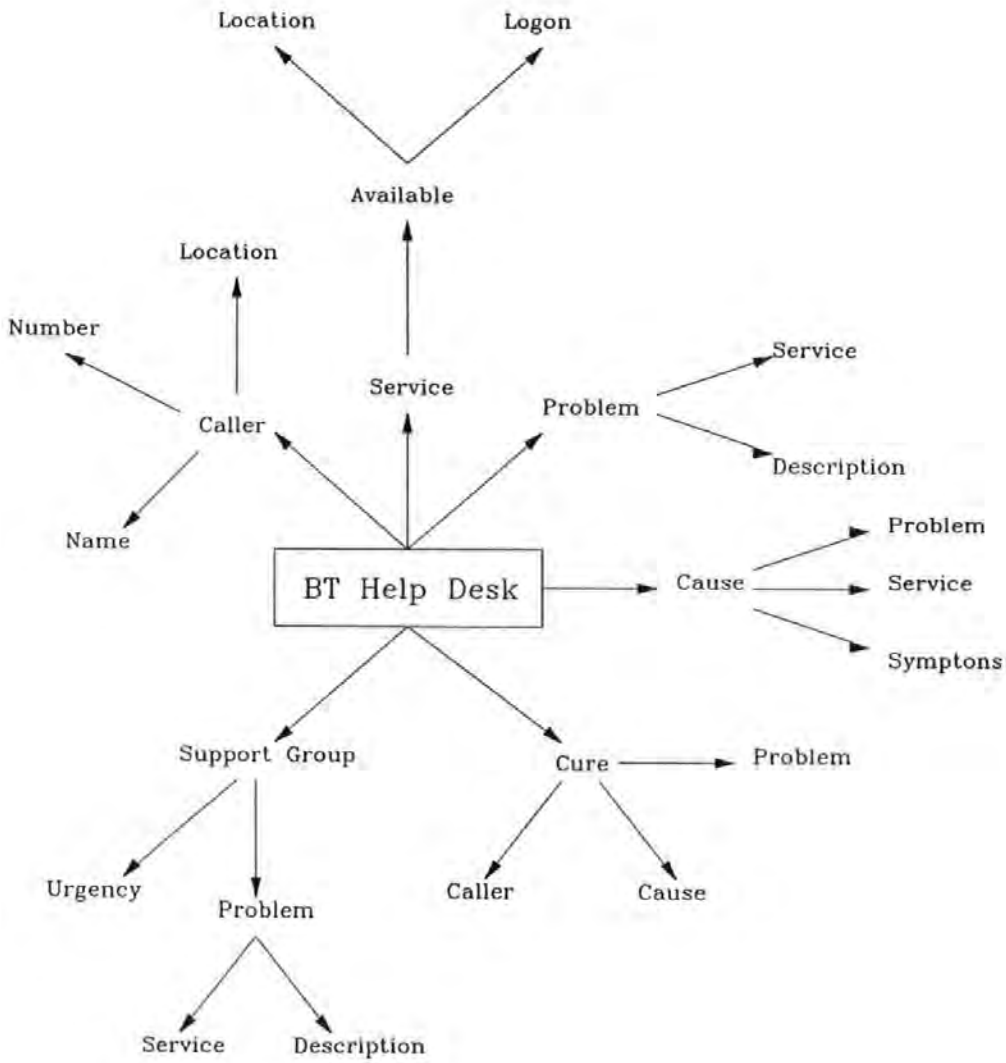


Figure 27 Free Engine Illustration (BT Fault Helpdesk)

Having represented the goals and sub-goal in this form the expert system inference mechanism has next to be implemented.

Prolog allows for two basic methods of inference:

forward chaining where rules are of the form:

If A THEN B & C;

(where the conditions can be conclusions, data items, or actions),

backward chaining where rules are of the form:

A TRUE IF B & C.

4.7 Conflict Resolution

During one cycle of the production system, it is most likely a common occurrence for more than one rule to be selected for use: one rule is chosen from among this 'conflict set' by a process of 'conflict resolution'. Different expert systems use different methods of conflict resolution including:

choosing the first rule that is matched;

choosing the highest priority rule (priority can be defined in different ways);

choosing a rule arbitrarily;

letting all the applicable rules fire.

Most expert systems will be restricted to one or two conflict resolution strategies. An increasing number of conflict strategies makes control over the consultation increasingly difficult to maintain. If only one conflict resolution strategy is offered, it is likely to be using of rules based on the numerical sequencing of rules in the knowledge base.

The following conflict 'rules' are traditionally considered the best: forward chaining rules should only be used when all their premises are true and should complete all their actions before the next cycle begins; global control strategies tend to produce a counter-intuitive order of questioning and it is necessary to control this effect; use of rules should be made conditional on the consequence of other rules.

4.8 Control

An expert system that has oversight of a large body of knowledge is not normally allowed to search blindly through all possible combinations of facts and conditions, synthesising new knowledge and hoping to arrive at a useful conclusion. A control mechanism, therefore, maps-out likely strategies and organises the overall effort. It evaluates the cost of pursuing a particular line of thought in terms of time, memory, suitability, etc.

CHAPTER 5

A NEW APPROACH TO NETWORK DESIGN

5.1 Introduction

Traditional network design methods assume that it is possible to generate not only the traffic sources and sinks but also communities of interest as well. This data is then presented in the form of a traffic matrix for its subsequent manipulation. This is not the norm: practical network design problems meet by BT Consultants have shown that such data are not generally available. To this end, details are given of three new algorithms which have been developed by the author, called:

Heuristic Demand Algorithm;

Node Location Algorithm;

Link Networking Algorithm;

that are able to overcome problems encountered in circumstances where there is insufficient information available for their solution using traditional methods.

The chapter concludes with an analysis of the impact of uncertainty on the resultant design.

5.2 A New Approach to Traffic Matrix Design

5.2.1 Estimating the Input Data

Opinion surveys are particularly suited to 'green field' situations where the basic telecommunication infrastructure has yet to be developed. In the heuristic demand algorithm, it will be shown how they are used in determining, in part, the customer demand in new, and estimating suppressed demand in existing, developments. In keeping with the objective of time series analysis, by comparing the growth of telecommunications in 'mature' scenarios over time, it is possible to predict the growth of demand in new telecommunication markets. In the heuristic demand algorithm, it will be shown that by the comparing the growth of telecommunication when measured against various economic indices the resultant correlations can be used in the determination of demand. Again, in keeping with the objective of cross sectional analysis, the heuristic demand algorithm will use economic data from many countries to predict demand in new situations. Call logging is not further considered as it presupposes that a network is available to monitor, which is not the case in 'green field' situations.

5.2.2 Heuristic Demand Model

Figure 28 outlines a model that was developed for calculating demand in the absence of call logging. The model describes a method used to build a traffic matrix based upon a set of heuristics. For example, telephone bills, site details, business operating environment and numbers of employees can be used to develop a profile of the total numbers of customers and the likely traffic that each generates. This data can also

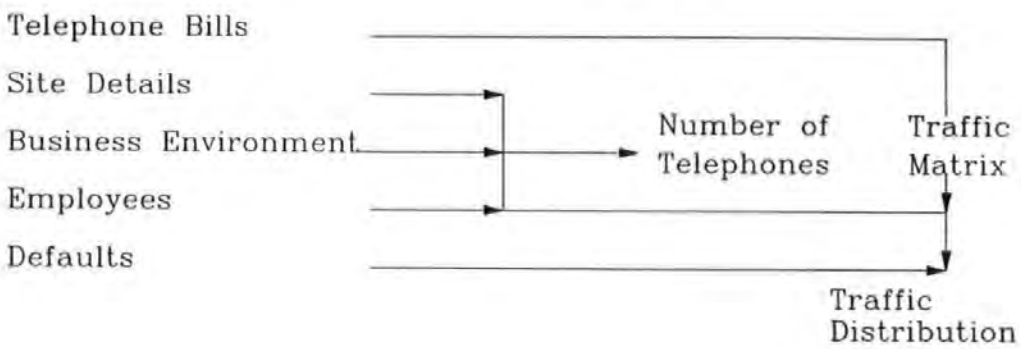


Figure 28 Heuristic Demand Model

be used to discover communities of interest. When used in conjunction with a set of default values to cover for unknowns, it is possible to generate a complete traffic matrix.

If the total number of telephones to be provided is unknown, it is possible to estimate this number based upon a site by site basis. One method, a heuristic found from the numerous consultancies conducted by BT Consultancy, is as follows.

Firstly, density functions are used where average telephones per square metre are available in a 'look-up' table. These averages are grouped under generic headings such as manufacturing, finance, remote office, HQ office, shop, etc.

Secondly, the synergy between company turnover and demand for telephone service will be shown in the following macro-analysis, in particular, that Gross Domestic Product (GDP) is directly related to telephone demand. In equating turnover of companies that comprise the 'Service Sector' segment of GDP, coarser but indicative results can be obtained.

Thirdly, analysis of employees' roles is necessary. For example, based upon BT Consultancy studies, one assumption is that between seven and seventeen per cent of the work force require telephone service; the percentage allocation dependent upon the function of the site under investigation. (This applies to the UK. However, the author has found that in countries which are culturally dependent upon the telephone, e.g. Japan and Hong Kong, the maximum rises to 25%, see appendix 2) The premise for the calculation is that executives demands will be greater than that of white collar workers, which in turn will be greater than that of blue collar workers.

Fourthly, once the above calculations are completed, it is necessary to apply a contention ratio that equates customers to network access lines. As a heuristic, a 10:1 contention is used for educational requirements, a 6:1 for business establishments and a 8:1 for groupings of office workers.

Finally, calling rates are applied whose value is dependant upon the class of customer. This traffic is then entered into the n*n traffic matrix. At this stage it is likely that the matrix is incomplete so that, if any vacancies exist in the matrix, the value from the diagonally opposite cell is copied, i.e. copy traffic from cell ij to cell ji, where ij represents the amount of traffic originating at node i and destined for node j. Thus making the assumption that the number of calls originating to a known location equals the calls received from the same location. This method of generating the source to destination data as a basis for network design is somewhat crude and, except for the simplest of network designs, not recommended. Under these conditions it is better to use the node location and link networking algorithms detailed later in this chapter.

The above heuristics are detailed in subsequent sections of this chapter.

5.2.3 The Heuristic Demand Algorithm

5.2.3.1 Number of Customers (Macro Analysis)

The most widely used heuristic method of determining customer demand derives from comparisons of availability and usage of telephone service with various measures of

national economic activity, such as GDP. This macro-economics (top-down analysis) evidence sets the scene for the more detailed micro analysis (bottom-up analysis).

By using the relationship with GDP, telecommunication infrastructure is viewed as a resource for a productive process just like petroleum or electricity. Consequently, most economic empirical work has assessed the effect of telecommunications at the macro-economics or country level in one of two ways.

The first, commonly called statistical correlation or regression analysis, consists of specifying a macro-economic country-level model, often with only one equation embodying supposedly causal relations, and then estimating the parameters of that model from statistical data. Such data includes the use of telecommunications and indicators of the level of country-wide economic activity. This approach includes both cross-sectional analysis, in which the variables are compared for different countries and regions at a single time, and time-series analysis, in which the variables of a single country are traced over time.

The second, referred to as structural economic analysis, focuses on the structure of the economy as revealed by the levels of activity in its different sectors, e.g. agriculture, manufacturing, services, etc. This approach relies primarily on input-output analysis and seeks to describe an economy in terms of stable coefficients relating the outputs of particular sectors to their requirements for inputs; one such input being the use of telecommunication services.

The CCITT [88] have identified the strong correlation between telephone density (the number of telephones per 100 persons) and what is called the 'wealth of nations'. The

equation representing this relation has been formulated many times using different groups of countries, and different periods, as well as per capita GDP or related indicators as proxy measures for the wealth of a country.

The CCITT used cross-sectional data from thirty industrial and developing countries to examine the correlation between the density of telephone access lines d and GDP per capita g. A scatter figure using logarithmic scales for both variables showed that most of the data clustered along a straight line, which can be represented by the equation

$$d = ag^b$$

or equivalently

$$\log d = b \log g + \log a$$

where log a is the intercept and b is the slope, see figure 29. The size of the parameter b indicates the rate at which telephone density varies relative to GDP increase together. If b is greater than one, then the higher the GDP, the faster d increases as GDP increases. The parameters log a and b were estimated by least-squares regression of log d and log g giving the following result:

$$\log d = 1.44 \log g - 3.0932$$

High correlation coefficient was obtained (0.92), indicating a strong cross-country relation between the variables.

Since b was greater than one, the rate at which telephone density increases relative to per capita GDP seems to be higher among the higher income countries than among the less-developed ones. To the extent that a cross section of countries indicates the

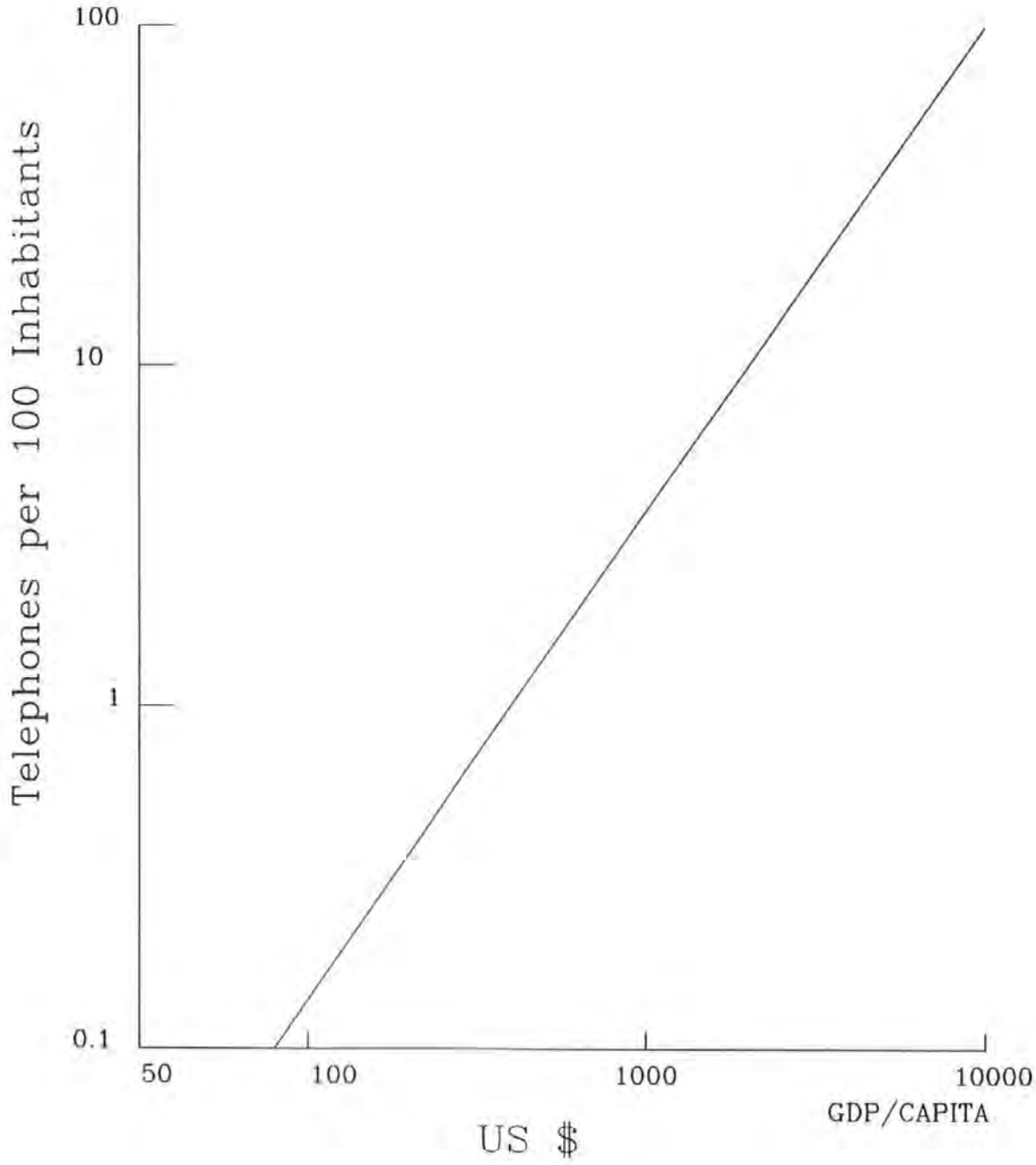


Figure 29 Cross-Sectional Analysis
(Telephone Density and GDP)

way in which telephone density may vary with per capita GDP through time in any one country, it appears that the more developed a country becomes, the faster its telephone service tends to grow relative to the rest of the economy.

The CCITT recommended in their GAS-5 handbook [88] that the slope of the fitted access lines be used to forecast both demand for, and supply of, telephones, in terms of expected growth of a countries GDP and its population, and also to forecast initial levels of demand and supply. This recommendation did not explicitly claim that the observed cross-country correlation implied a causal link between telephone provision and the growth of GDP, although it was suggested elsewhere in the handbook that such relations can indeed indicate the benefits of telecommunications.

An alternative to examining a cross section of many countries at a given time is to examine the density of telephone access lines, for a single country, as its per capita GDP increases through time. Figure 30 shows an exercise of this type for Sweden, reported by the CCITT in its GAS-5 handbook.

The Swedish data show two different trends, which represent different rates of exponential growth of telephone density in relation to the corresponding growth rate of 'per capita GDP'. These tests can be interpreted as first a trend reflecting the period in which telephone service was introduced. The second trend shows the more gradual process of connecting large proportions of the population to the system. The parameters of the equations representing the data estimated by least-squares regression of log d and log g separately for the two periods, were:

$$\log d = 3.1935 \log g - 10.4106$$
$$\log d = 1.5476 \log g - 4.64450$$

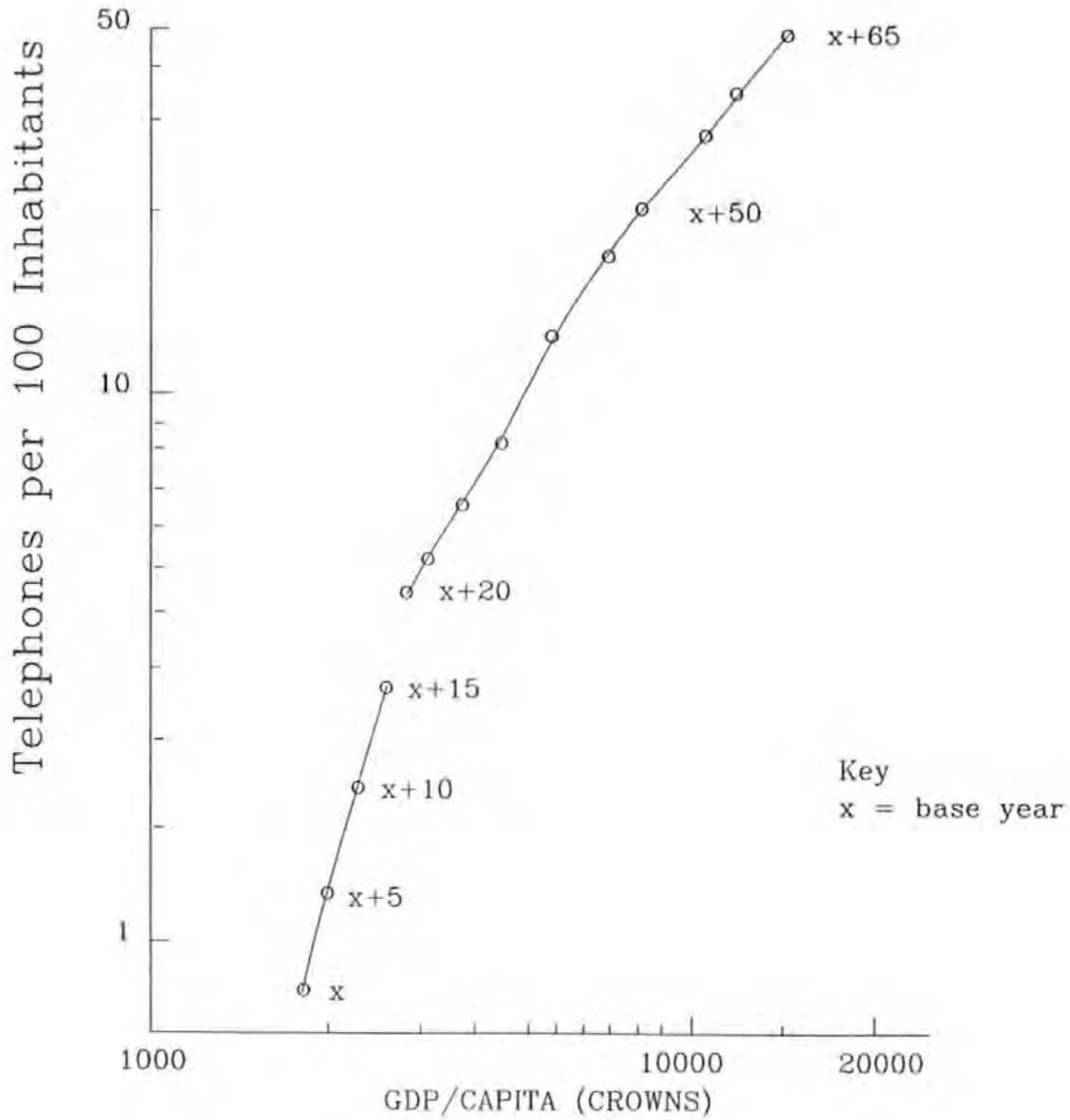


Figure 30 Time Series Analysis
(Telephone Density–Sweden)

High coefficients of correlation were obtained, indicating a strong relation between the variables. As in the cross-sectional correlation, the curves represented by these equations are concave upward when drawn with linear scales.

The second coefficient of log g, (1.5476), is similar in size to those for the cross-sectional models discussed previously (1.44), and in this case implied for Sweden that as GDP increased, telephone density increased more rapidly.

Several other correlations involving GDP, have been tested. One exercise [88] made use of a so-called telecommunication 'utilisation factor', defined as the number of telephone connections per US $100,000 of GDP. Examination of the utilisation factor for fifty-five economies had high utilisation factors (ten or more for the United States, Canada, and most of Western Europe), lower factors were found among developing economies with a strong industrial sector (five to eight for Brazil, the Republic of Korea, Mexico, Singapore, and Taiwan). The lowest utilisation factors (typically one to three) tended to be associated with the developing economies. This analysis simply suggests that increasing growth in telecommunication services correlated with economic grow. They do, however, suggest that the observed disparities in utilisation factors can be used as a general argument for developing countries to raise their priorities for investment in the telecommunications sector. This implies that a link exists between telecommunication investment and GDP. Analysis of the International Monetary Fund (IMF) [89] and CCITT [90] statistics show that most developing countries have 1% of GDP invested in telecommunication infrastructure whilst those of Western Europe and North America approximately 2%: the IMF forecast that investment in telecommunication infrastructure in western countries will increase to 7% of GDP by the year 2000. The link between telecommunication

investment and number of customers of a telecommunication network is more tenuous; one can calculate the cost of one connection between two customers including all link and nodal equipment forming the denominator of such a calculating formula.

The author has conducted other exercises based on statistical correlation to improve the representation of economic and telecommunication activities. For example, a strong relation between the number of telephone customers and the total consumption of electricity, as well as between the ratio of telephone to electricity customers have been observed. None have been proven to be reliable in situations outside the country from whose statistical data was used.

In summary, by a process of interpolation and regression analysis, the above formula can be encapsulated into a usable heuristic form as follows:

$$\text{Telephones (per 100 population)} = ((\text{GDP(in US\$)} * 0.015) - 4).$$

5.2.3.2 Number of Customers (Micro Analysis)

The top-down, macro economic, analysis is useful in identifying the total requirement for telephone service; it can also be used to check bottom-up, micro analysis, whereby the resultant value of summation of the micro data should tend towards the value calculated by macro economic means.

One micro analytical method, developed by the author whilst conducting a significant number of network designs for BT, assumes that the demand for telephone service is correlated primarily with buildings and their usage. Demand arises from the

following two sources: the first is new developments (office buildings, condominiums or housing estates): including projects under construction, projects that have planning permission and assessing the likelihood of the their construction and those unplanned but there exists a likelihood of their construction. Here, telephone access lines are traditionally provided by the building or housing estate developers: additional lines may be required by occupants.

A second micro analytical method attempts to quantity 'suppressed' demand. The following heuristic approach, developed by the author, can be employed to quantifying the number of telephones in a micro economic way.

First it is necessary to determine the total area of buildings and their probable mix of occupants by type of business. Then determine the likely requirement of telephones, per square metre, per type of business. From the product of the above, deduct the number of direct access lines and extensions (in so far as they are considered substitutes for direct external lines) to calculate the suppressed demand for services. Now apply a contention ratio to estimate the number of access lines for each building.

Buildings will have different telephone densities dependant upon their utilisation. For example, government offices are similar to commercial buildings but with a smaller number of lines required per square metre. Residential properties will require the application of penetration factors on building types and locations. In the worst case each house or flat in a high rise building will require one line. Hotels are a special case since they have a large number of rooms and telephone line demand is not in the same proportion to the number of square metres or space in the building as for commercial and residential. In terms of telephone usage, hospitals are similar to

hotels but with a smaller number of lines required per square metre. With shopping centres, it is assumed that each shop will require one line and that each department store will require a number in proportion to its selling area.

The telephone densities either the number of telephones per square metre, or telephones per pound sterling (£) of turnover or other relationships, are best obtained from a sample survey conducted amongst a mix of business types, such as architects, estate agents, developers and telephone companies.

Estimates of demand for each major building or housing estate can then be based upon: estimates by occupiers; actual use per square metre per employee in 'mature' commercial areas, (on the assumption that the newer commercial areas will eventually reach those levels); and comparison with actual use in other networks per square metre, per employee, by type of company. One such formula is:

$$\text{Number of telephones per Unit Area (Business Function)} = \frac{\text{Total Number of Telephones}}{\text{Total Area}} \text{(Business Function)}$$

5.2.4 Traffic Generated

Having calculated the number and distribution of customers and hence telephones that will use the network, it is now necessary to translate these quantities into traffic values so that the required link and node equipment can be dimensioned.

5.2.4.1 Traffic Generated (Macro Analysis)

Correlation analysis has been used to explain variations in telecommunication traffic across countries and through time. For instance, in the GAS-5 handbook [88], the CCITT reported on several traffic studies based on time-series data for individual countries including exercises consisting of correlation between telephone calling rates and GDP, GDP per capita, value of imports, value of exports and other measures of economic activity. Results were mixed. The CCITT concluded that, with the data available, local telephone call traffic could not be shown to be influenced by the normal course of economic development.

The author has shown, in his design of various international networks, that long distance telephone traffic increased at roughly double the rate of increases in real GDP. However, since wide variations occurred among countries and from one year to the next, this rate could not be used as a universal benchmark. Not surprisingly, international telephone traffic through time was most closely associated with changes in the volume of international trade.

Regarding cross-sectional exercises, the author has studied forty-six industrial and developing countries and analysed the correlation between international telephone traffic and twenty economic indicators specified as dependent variables. The numbers of tourists per year, GDP per capita, and size of the country population were found to be the main variables explaining international telephone traffic among countries. Whilst all three were positively associated with telecommunication traffic no general formula could be developed.

The UK government has shown that the telecommunication resource is mostly consumed by the service sector industries, whereas the manufacturing sector group generally follows in second place [91], thus, as a heuristic, it is reasonable to say that 'local' traffic will roughly double with a doubling of the service sector contribution to GDP [98].

5.2.4.2 Traffic Generated (Micro Analysis)

Macro-analysis proves to be an unreliable method of attempting to determine the traffic for designing networks, micro-analysis offers an alternative route to solving the problem. There needs to be a way to convert the total number of telephones into the total traffic generated at a site-by-site base; this need is satisfied by using various estimation techniques, each having an intrinsic uncertainty of varying degrees. These techniques include:

penetration factors, where different types of building are representative of the potential customer they house and are assumed to have an approximate level of telephone penetration at an average calling rate;

telephone bills, where call bills can be translated into traffic (erlangs or call connect seconds equivalent);

job functions, where stereo-typical functions have a traffic profile; or customer status within a company or society where again they can be associated with average traffic profiles;

traffic per 'pound sterling' of company turnover, and numerous default values.

Of the above, the most commonly available data are telephone bills. It is possible to convert the telephone call-charge bill into traffic if assumptions are made. The following assumptions have been found by the author to produce good correlation with actual measured traffic:

the charge comprises 90% of local 10% of trunk calls;
trunk calls are made up a middle tariff range (e.g. BTs 'B' band);
busy hour traffic is one fifth the day traffic.

This allows the following heuristic to be used:

$$\text{Busy Hour Traffic} = \frac{(12 \text{ month bill})}{(\text{Unit Charge} * 4 * 365 * 5 * 60)}$$

From this the average BH traffic per telephone is:

$$\text{Average BH traffic} = \frac{\text{BH traffic}}{\text{No. of telephones on bill}}$$

The busy hour call attempts (BHCA) needs to make allowances for 20% of calls failing and average call duration of three minutes:

$$\text{BHCA} = \frac{(\text{average BH traffic} * 60)}{3}(1.2 * No.Customers)$$

If a check on the traffic estimated above is required, then further estimates of traffic calculated as a function of both business turnover and function can be used One heuristic found by BT consultants is that all customers with twice the national average income generate 0.3 erlangs of traffic: three times the average, 0.5 erlangs of traffic. BT consultants also found that for each site a minimum default of 10 access lines at 0.02 erlangs per connection is set with a traffic split of 80% local and 20% trunk being assumed.

5.3 A New Approach to Nodal Network Design

The author has, through the process of conducting a number of consultancies **[92 to 98],** been able to design a new process of converting traffic from a point source matrix to nodal network components.

Traditional methods used in generating the traffic matrix have assumed that it has been possible to identify traffic distribution. It is usual, under such circumstances to use crude approximates of percentage of local, adjacent and trunk calls, to produce a source and destination matrix. To ignore this problem, it is necessary to consider other techniques. First, the designer needs to decide how many nodes are needed to support the total traffic of the area under investigation. This is achieved by first calculating the total traffic the area generates by the expression:

$$\text{Total Traffic(Current)} = \frac{(((\text{GDP(in US Dollars)} * 0.015) - 4) * \text{Population} * 0.05)}{100};$$

assuming that the average calling rate per customer is 0.05 erlangs per connection in the busy hour.

To forecast the traffic for the end of the design period, the following relationship is appropriate:

$$\text{Total Traffic(Planned)} = \text{Total Traffic(Current)} * \frac{\text{GDP(Service Sector) Forecast at year of plan}}{\text{GDP(Service Sector) current}}$$

To calculate the number of nodes necessary to support this level of traffic, it is necessary to know whether the solution is a private or public switched network. This is because the traffic capacities of the two types of nodes are somewhat different. Private nodes are essentially building based and relatively small, their overall total capacities are typically about BH 200 erlangs. Public nodes, whose collection areas can be as large as 12.6 square kilometres, have capacities of the order BH 2000 erlangs.

Hence the total number of nodes required for the area under investigation is approximately:

$$\approx \frac{\text{Traffic (Planned)}}{\text{Node Capacity}}$$

5.3.1 Node Locations and Traffic Collection Areas

Having now calculated the number of nodes required to support the whole of the traffic, it is now necessary to decide where best these nodes should be placed.

A node's natural location is where the total cost of the local access network is at a minimum. This situation occurs when the total weighted aggregate length of the access network is at a minimum.

The general problem to be solved is to derive simultaneously the optimum node locations and their service areas, based upon the optimum number of nodes found by the macro-economic country data.

With this in mind, a matrix erosion process is able to identify, first the optimum locations and second the optimum service areas, of these nodes. In essence, the process is one of identifying which of the 250*250 metre squares comprising the traffic matrix are peaks, in a three dimensional representation of the traffic against geographic location, these peaks forming the node locations. If there are fewer peaks than the number of nodes required then the maximum traffic level is gradually reduced, erlang by erlang, until the number of new peaks produced meets the number of node criteria. The optimum service area of each of the nodes is produced by the allocation of each square to its nearest node, measured as radial distance.

Consider, a three dimensional plot of traffic, on the z axis plus x, y grid reference as shown in figure 31.

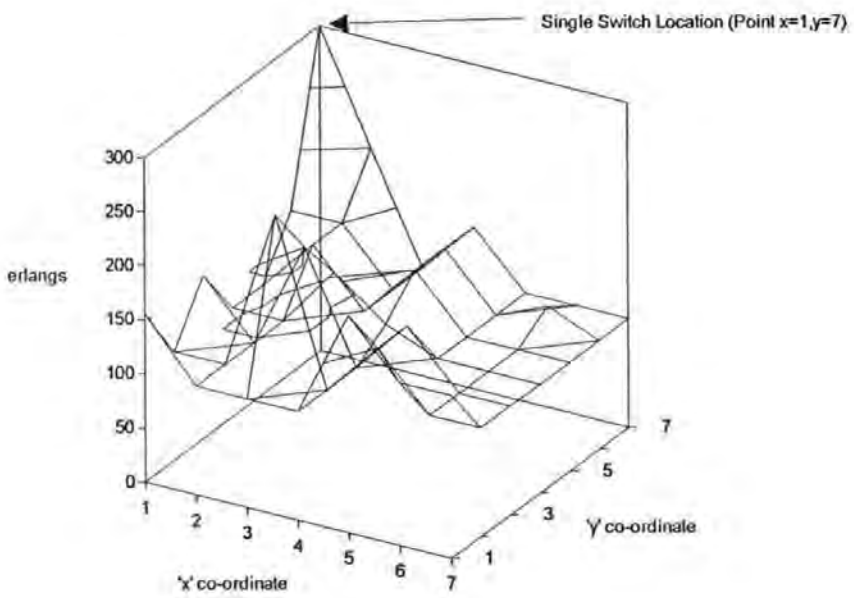


Figure 31 Three Dimensional Plot of Traffic Matrix

The peak on the plot, x=1 and y=7, indicates the location for a single node requirement. However, if the total number of nodes required for the area under investigation is two (found from the macro-analysis), the peak traffic value in the point source matrix needs to be reduced until two maximum and equal traffic levels result. Figure 32 shows the result of this iteration (e.g. erosion from 300 to 250 erlangs).

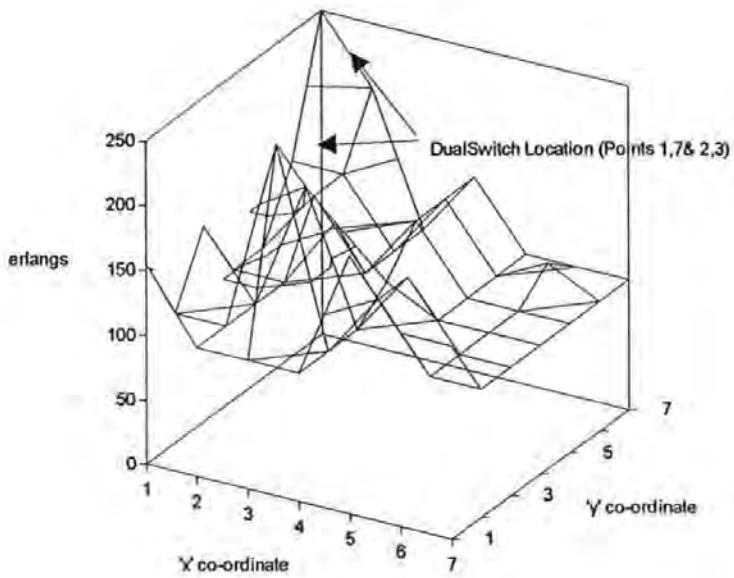


Figure 32 Plot of Eroded Traffic Matrix

This erosion process continues until the location of the required number of nodes is found. Figure 33 shows the results for three nodes (e.g. 200 erlangs).

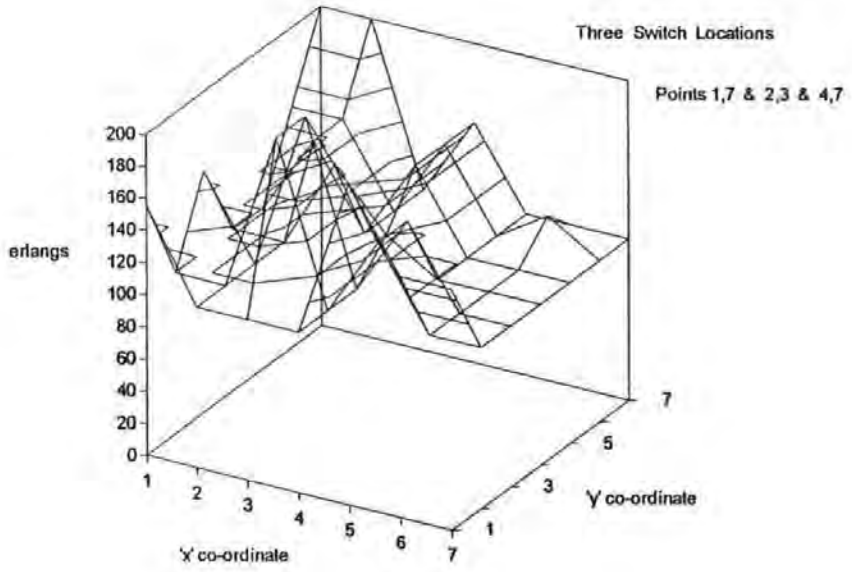


Figure 33 Plot of Further Eroded Traffic Matrix

5.3.2.1 The Nodes

Within the 250*250 metre square, in areas with simple cable and duct layouts, the optimum node centres will be located near one or two intersection points of existing and potential cable routes. In congested areas, and where the road network is complex, several such intersection points may exist, each which may be a viable location for node centres. In both cases it is necessary to carry out a detailed study of the costs involved to determine the best practical location. Any one of the following may influence the final location of the nodes:

availability of site;

price of available site;

local building regulations.

These factors need be further influenced by the following node types.

5.3.2.2 Transit Nodes

A heuristic approach is the simplest way in finding the practical transit centre location. The following algorithm is one way in which such a study might be conducted.

On the scale map of the area under investigation, which has been divided into 250*250 metre squares on which existing cable routes should be identified, the customer forecast is distributed into the appropriate squares. As before, the optimum square to hold the transit is found at the intersection of the mid traffic value of the rows and columns, i.e. where the difference between the total traffic to the left and to the right, and similarly above and below, is a minimum. If the square contains a major intersection then this remains the practical location, else, move to the nearest contiguous square that has the greatest density of intersections.

5.3.2.3 Local Nodes

The process of matrix erosion has already placed the local node within the square with the greatest traffic levels. It is therefore unlikely that the chosen location will have to change. It is essential that there is an adequate intersection point on which to place the node. The database that holds the percentage primary, secondary and tertiary land utilisation needs to be enhanced to be able to describe adequately the road, and hence duct, network.

In addition, the default node size has previously been taken as 10,000 access lines. However, a survey of BT nodes has shown that those located within a City Core area

support approximately 22,000 access lines; within Suburbia 7,000 access lines and within Fringe areas 1,000 access lines. Indeed, Fringe area nodes can, in practice, be implemented as Remote Peripheral Units parented upon a suburban node.

5.3.2.4 Node Collection Areas

Previously, it was shown how it was possible, by a process of matrix erosion, to define the node collection areas. In practice, it is necessary to modify the collection areas to reflect cable routeing constraints. An example of a routeing constraint is an area divided by a river. The river would, unless there were many bridges, provide for the natural boundary of the collection area.

If the position exists were many bridges are available then the traffic collection area needs to be divided into a number of 'obstacle zones'. Distances between 250*250 metre squares within the same obstacle zone are expressed as rectangular. Distances between squares in different obstacle zones are calculated by passing through the routeing point, i.e. the bridge, or combination of routeing points, again using rectangular distances within each obstacle zone.

To aid the pragmatic approach the land utilisation matrix needs to be enhanced with the following information:

description and location of the obstacle;

location of the routeing points.

5.3.3 The Node Location Algorithm

The above examples indicate the algorithmic process used in determining the location of both the local and transit nodes. In practice the matrix erosion process, when applied in the general case, is implemented in the following way.

Given that:

the number of nodes required to service the area is known from the macro-economic analysis;

the maximum radius of each node service area is 2 kms., i.e. 4 kms. worst case access line, thus dictating that the maximum number of 250*250 metre squares that can be collected by a node, in each quadrant, is thirty-six, (this equates to a 45^{o} vector distance of approximately 2 kms). This means that the maximum traffic collecting area of a node is 144 squares, i.e. 9 square kms.; and

only whole 250*250 metre squares can be allocated;

The algorithm operates as follows.

(a) Round-up all the squares' traffic values, that form the 'point source traffic matrix', to the nearest 10 erlangs.

(b) Find the maximum traffic value, T_{max}.

(c) Search array, starting from square '1,1' and checking each unmarked square, find first occurrence of $T_{max.}$; mark square as potential node site and mark all adjacent sites, i.e. for square x,y mark squares '(x-1 to x+1),(y-1 to y+1)'.

(d) Sum $T_{((x-1 \text{ to } x+1),(y-1 \text{ to } y+1))}$ as $T^*_{x,y}$.

(e) Search every other square, each that has a traffic value = T_{max} marking and summing adjacent traffics as in (c) and (d).

(f) When search is complete, three conditions are possible:

i) number of marked squares equals the number of nodes required,

Go to (I).

ii) number of marked squares is less than number of nodes required,

Go to (g).

iii) number of marked squares is greater than number of nodes required,

Go to (h).

(g) Reduce T_{max} by 10 erlangs and repeat (c) to (f).

(h) Select the squares equal to the number of nodes required in descending values of T^*. Mark the parent square as node locations; i.e. parent of $T^*_{x,y} = T_{x,y}$.

(i) Proceed with identification of the service collect areas. Taking each square in turn find the nearest node, i.e. $\min(|x-A|^2 + |y-B|^2)^{0.5}$, where x,y are the square co-ordinates and A,B the node co-ordinates. When two nodes are

equidistant from the square being evaluated, allocation is made to the node with the lowest current cumulative number of collected squares.

(j) Next it is necessary to check that the node is acceptable for the square, '(x-6 to x+6),(y-6 to y+6)', i.e. within the 2 km range, allocate and return to (i), else find next nearest node and return to (i),etc. If no node proves acceptable the square is marked as node site, return to (i).

The above algorithm needs to be modified if transit working is a prerequisite of the design. The process now becomes:

(a*) Round-up all the squares' traffic values, that form the 'point source traffic matrix', to the nearest 10 erlangs.

(b*) Find the maximum traffic value, T_{max}.

(c*) Find the square that is the traffic centre of the total area under consideration. The location of this square is at a point where the difference between the total traffic to the left and to the right of the square, and similarly above and below the transit, is a minimum. Mark square as a definite node and transit site and mark all adjacent sites, i.e. for square x,y mark squares '(x-1 to x+1),(y-1 to y+1)'.

(d*) Search array, starting from square '1,1' and checking each unmarked square, find first occurrence of $T_{max.}$; mark square as potential node site and mark all adjacent sites, i.e. for square x,y mark squares '(x-1 to x+1),(y-1 to y+1)'.

(e*) Sum $T_{((x-1 \text{ to } x+1),(y-1 \text{ to } y+1))}$ as $T^*_{x,y}$.

(f*) Search every other square, marking each that has a traffic value = T_{max} and summing as in (d*) and (e*).

(g*) When search is complete, three conditions are possible:

i) number of marked squares equals the number of nodes required,
Go to (j*);

ii) number of marked squares is less than number of nodes required,
Go to (h*);

iii) number of marked squares is greater than number of nodes required,
Go to (i*).

(h*) Reduce T_{max} by 10 erlangs and repeat (d*) to (g*).

(i*) In addition to the square found as transit location, select the squares equal to the number of nodes, less one, required in descending values of T*. Mark the parent square as node locations; i.e. parent of $T^*_{x,y} = T_{x,y}$.

(j*) Proceed with identification of the service collect areas. Taking each square in turn find the nearest node, i.e. $\min(|x-A|^2 + |y-B|^2)^{0.5}$, where x,y are the square co-ordinates and A,B the node co-ordinates. When two nodes are equidistant from the square allocation being evaluated, is made to the node with the lowest current cumulative number of collected squares.

(k*) Check that the node is acceptable for the square, '(x-6 to x+6),(y-6 to y+6)', i.e. within the 2 km range, then allocate and return to (j*), else find next nearest node and return to (j*),etc. If no node proves acceptable the square is marked as node site, and return to (j*).

A subsequent stage in the analysis process needs to decide whether to open each of the nodes. This is based upon the comparison of the cost nodal against links based upon the laborious task of calculating the total pair-kms. required for each alternative. Naturally, other factors influencing the final decision, including the cost of opening a new building, and duct route that will only be used for part of their useful life.

5.3.3.1 Node Location in Action - Examples

This heuristic method is best illustrated by way of a simple example.

Consider two urban 'areas', each having an area of 9 square kms., and each is to be served by differing number of nodes. On the basis of the relationship of Gross Regional Product (GRP), telephone penetration and digital node capacity, the first urban 'area' requires just three local nodes and the second urban 'area', five local nodes.

The example is developed by the addition of a requirement that for the second 'area' also requires a transit node, necessary because it needs to intercommunicate with other urban areas.

For each urban area we first assume that the point source traffic matrix has been developed.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

5.3.3.1.1 Urban area requiring three local nodes

Figures 34 through 43 show how using the process of 'matrix erosion' and how the collection areas of the three nodes are developed. First, the three nodes are located at the points of greatest traffic density, A1, D4 and F1 in this example. The first 'erosion pass', takes the maximum point source traffic level down from 100 to 90 erlangs; node A1 'collects' the areas B1 and A2; F1 collects F2; and D4 collects C3, D5, A5, B6 and F5, all of which are radially closer to D4 than the other two nodes.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 34 First Erosion Pass

A second erosion pass takes the maximum traffic level down from 90 to 80 erlangs. Using radial distance as a rule-of-collection, the node located at D4 collects the further areas of D2, F4, E5, C6 and A4; node A1 collects B3.

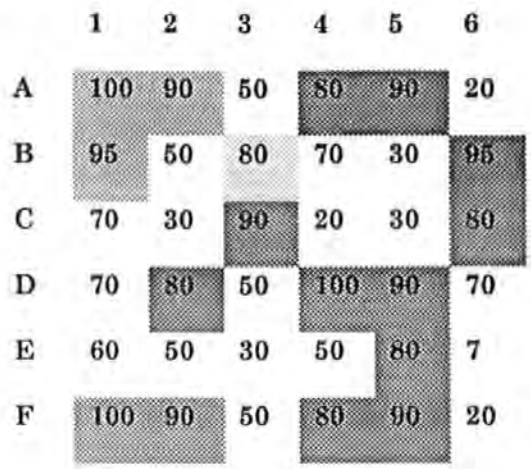


Figure 35 Second Erosion Pass

The last two square allocations were arbitrary as the areas A4 and B3 are equidistant to both nodes A1 and D4. The heuristic is that under these circumstances the allocation is shared.

The process of erosion and area collection continues until all squares are allocated to a node. In this example, the final node boundaries are shown in figure 36.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 36 Third Erosion Pass

5.3.3.1.2 Urban area requiring five local nodes

The matrix erosion and traffic collection processes for five nodes are the same as above. At the first stage of erosion only three of the nodes can be allocated; A1, D4 and F1.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 37 First Erosion Pass

Erosion of the maximum traffic values in the point source matrix to 95 erlangs identifies a further node location at B6. In addition, B1 is collected by node A1.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 38 Second Erosion Pass

The next erosion stage takes the point source matrix down to 90 erlangs. Area A2 is collected by node A1; A5 by node B6, C3 and D5 by node D4 and F2 by node F1. However, a location for the fifth node is established, F5.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 39 Third Erosion Pass

The erosion process continues as above until all areas have been allocated as shown in figure 40.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 40 Fourth Erosion Pass

5.3.3.1.3 Urban area requiring three local nodes and a transit

Before determining the location of the nodes and their collection areas, if it is a design requirement that a transit node is necessary or will be used in the future, it is first necessary to determine the location of that transit node. The location of this node is at a point where the difference between the total traffic to the left and to the right of the transit, and similarly above and below the transit, is a minimum. This condition provides an easy method for finding the optimum location.

In this example the centre of the matrix is at half the total cumulative traffic, i.e. 1168 erlangs. The matrix below shows the totals of the 'x' and 'y' traffics.

	1	2	3	4	5	6	TOTAL	CUM
A	100	90	50	80	90	20	430	430
B	95	50	80	70	30	95	420	850
C	70	30	90	20	30	80	320	1170
D	70	80	50	100	90	70	460	1630
E	60	50	30	50	80	7	277	1907
F	100	90	50	80	90	20	430	2337
TOTAL	495	390	350	400	410	292	2337	
CUM	495	885	1235	1635	2045	2337		

Figure 41 Transit Location

The centre of the matrix is located at a (x,y) traffic of 1168 erlangs, i.e. square C3.

The process of finding the node location(s) is now the same as the examples without transits except now one of the nodes has now to be located at the transit site, C3, this is irrespective of the squares intrinsic traffic density. The following figure shows the location of the three nodes. No erosion is necessary as there are three locations with 100 erlangs besides the transit site. Square D4 is 'collected' by the transit site C3: A1 and F1 form the location of the other two node sites.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 42 First Erosion Pass

The erosion process continues until all areas are reduced to zero erlangs. The resultant is shown in figure 43. As the transit is located near the original node sites, the resultant service areas are similar, although this need not be the case.

	1	2	3	4	5	6
A	100	90	50	80	90	20
B	95	50	80	70	30	95
C	70	30	90	20	30	80
D	70	80	50	100	90	70
E	60	50	30	50	80	7
F	100	90	50	80	90	20

Figure 43 Final Erosion Pass

5.4 A New Approach to Link Network Design

The graph theory algorithms detailed in previous chapters require, as a premise, that routeing of calls are through both transits, i.e. parent and destination. However, two other routeing strategies are available to the network designer, routeing through the parent transit located in the originating area; routeing through the transit located in the destination area. Routeing disciplines do not in themselves produce significant economic changes, particularly when alternative routeing is used, as optimum dimensioning of high usage and overflow traffic routes compensates for possible mistakes in routeing arrangements. Operational and security requirements favour a simple routeing scheme, to be adopted in preference to that which is theoretically the optimum, without incurring significant cost penalties.

Examples of actual network designs show patterns take a particular form. The author has, through the process of conducting a number of consultancies **[92 to 98]**, been able to reveals the following:

NETWORK	NODES	LINKS	*FACTOR	TYPE
BP OIL(OLD)	10	9	0.90	STAR
BP OIL(NEW)	18	17	0.94	STAR
BP DEVELOPMENT	12	13	1.08	STAR
BP CHEMICALS	25	25	1.00	STAR
BP INTERNATIONAL	11	14	1.27	STAR
BP GROUPNET	116	124	1.06	NET
REED INTERNATIONAL	61	60	0.98	STAR
PAKISTAN CELLULAR	8	10	1.25	NET
BT PRIVATE NETWORK	46	46	1.00	STAR
BT PUBLIC NETWORK	6830	9316	1.36	NET

These results indicate that as heuristics:

the number of links used in a minimum cost practical network design will not exceed either 1.36 or 1.27 the number-of-nodes for public and private networks receptively;

networks are not more than two levels deep;

nodes within the same transit group communicate by way of their parent transit whilst those in different transit groups communicate by way of their respective parent transits, and the links connecting them.

These three heuristics are used to simplify the graph algorithms when used for telecommunication network design.

5.4.1 The Link Networking Algorithm

The structures of telecommunication networks indicate that the Esau-Williams graph algorithm can be modified by boundary conditions that produce results favourable to both public and private network design and provide for a reduction of calculation complexity. For example, a typical network design is equivalent to a two level hierarchical-net. Two nodes within the same transit group communicates by way of their common transit node. Traffic from a node to one in another group is routed through the transit network, along a suitable link to the appropriate transit then down, through the service and supply network to the target node. In this way the transits collect and distribute traffic around the network. This structure is particularly effective. Much of the traffic generated within a node is likely to remain local to that node or one of its neighbouring nodes; the multiple star arrangements at

the lower level of the network deal with this well. The complete flexibility of the higher level, on the other hand, allows transits to be connected in the manner most appropriate to the characteristics and distribution of the traffic between them.

Thus as a first stage in the design, it is necessary to design a two level hierarchical-star. The second stage involves the adding of inter-transit routes to meet grade of service requirements. The results are a hierarchical-net giving the required network grade of service. The constraints are that the node i should not be more than two 'hops' from the centre and traffic capacities of arcs are not exceeded.

The Link Networking Algorithm constructs a constrained minimal spanning tree of a graph G with n nodes and weighted edges. The algorithm proceeds as follows:

a) construct a minimum spanning tree of the form 'Star' on each node (i=1 to n) in turn;

b) select that node acting as centre which gives the minimum cost 'Star' tree;

c) calculate cost T_{ij} for each pair (ij);

d) do while T_{ij} <> zero;

e) find the pair (i*,j*) such that T_{i*j*} is a maximum;

The constraints are that node i^* should not be more than two 'hops' from the centre and traffic capacity of links are not exceeded and the cost of the new link plus enhanced link to transit is less than $T_{i^*j^*}$.

f) If none of the constraints are violated by connecting i^* to j^* and disconnecting i^* from the centre, then do so; set $T_{i^*j^*}$ = zero and go to (e). Otherwise set $T_{i^*j^*}$ = zero and go to (e).

This gives a hierarchical-star configuration for nodes and their transit. A hierarchical-net is then constructed by the addition of links between level two star networks so that the overall GOS is met.

g) Calculate GOS_{xy} for each pair(x^*,y^*), such that $GOS_{x^*y^*}$ is a minimum.

h) Do while $GOS_{xy} <> 0$.

j) If the constraints are not violated by connecting x^* to y^*, then do so, and set $GOS_{x^*y^*} = 0$ and go to (g). Otherwise set $GOS_{x^*y^*} = 0$ and go to (g).

To dimension the links it is first necessary to define, and subsequently to calculate, the end-to-end network grade of service; since the network takes the form of either a hierarchical-star or hierarchical-net the worst-case grade of service can be found by adding the individual link losses.

Having determined the necessary link grade of service it now becomes necessary to determine the number of channels required to carry the traffic. As a heuristic, it is

possible to say that, on average, the BT tariff of 15 single channel 64 kbit/sec equates to one 30 channel 2.048 Mbit/sec. However, if the aim is to carry telephony and data on the same links, then it is not considered prudent to install more than 10 channels per link. This break point allows the simplification of the Erlangs B formula to give the following look-up table of channel capacities:

Traffic	Grade of Service	
erlangs	0.001	0.01
<0.1	2 channels	2 channels
0.2	3 channels	2 channels
0.3	4 channels	3 channels
0.4	4 channels	3 channels
0.5	5 channels	4 channels
0.6	5 channels	4 channels
0.7	5 channels	4 channels
0.8	6 channels	4 channels
0.9	6 channels	5 channels
1.0	6 channels	5 channels
1.5	7 channels	6 channels
2.0	8 channels	7 channels
2.5	9 channels	7 channels
3.0	1*2 Mbit/sec	8 channels
3.5	1*2 Mbit/sec	9 channels
4.0	1*2 Mbit/sec	1*2 Mbit/sec

Thus, the above shows as a simple heuristic that the break points for the provision of 2 Mbit/sec channels carrying voice and data are 3 erlangs (GOS = 1 call lost in 1000) and 4 erlangs (GOS = 1 call lost in 100). The first 30 channel group needs to be augmented at link traffics of 18 erlangs and 20 erlangs respectively by a further 30 channel group and an additional group at each further 20 erlangs until the traffic is greater than 90 erlangs then a linear approximation may be used with relatively small errors as shown below:

Traffic erlangs	Grade of Service 0.001	0.01
4<T <21.0	1*2 Mbit/sec	1*2 Mbit/sec
24.0	2*2 Mbit/sec	1*2 Mbit/sec
24.1	2*2 Mbit/sec	2*2 Mbit/sec
41.0	3*2 Mbit/sec	2*2 Mbit/sec
44.0	3*2 Mbit/sec	3*2 Mbit/sec
61.0	4*2 Mbit/sec	4*2 Mbit/sec
81.0	5*2 Mbit/sec	5*2 Mbit/sec
>90.0	Round-up((E+35)/30)	*2 Mbit/sec

[To convert from erlangs to hundred call seconds per hour (ccs/h), multiply the number of erlangs by 36.]

The cost of a node comprises a fixed cost covering the central processing unit, power supplies and common equipment plus the marginal cost of providing equipment for each digital channel terminating on the node. This cost can be expressed as Cs=At+Btd*Pd, where At is the common cost, Pd is the number of digital input/output ports entering or leaving the node, and Btd is the cost per digital input/output port at the node.

The overall cost of the node and link components of the network is thus given as follows:

Cn = (sum of all node costs + sum of all digital link costs). The aim of optimisation algorithms is usually to make Cn a minimum.

5.4.2 Routeing Optimisation

Berry **[99]** and Pratt **[100]** have developed a set of explicit formula for the dimensioning of links offered overflow traffic. This formula requires a rather complex

calculation, an interpolation of the results of plots of their formula and curve fitting to the interpolated data results in a heuristic the number of channels, n, for overflow traffic approximates to $n = (A[\frac{A6}{(9+12A)^{0.5}-3}]-1)$, where A is the traffic in erlangs and GOS is 1:100.

It is possible to simplify further their formula and define the following algorithm:
If traffic under investigation for the link is non-overflow type, and,

< 25 erlangs then send traffic on direct link to transit;

> 25 erlangs but < 50 erlangs put 25 erlangs on direct link and remainder on transit link;

> 50 erlangs but < 75 erlangs put 50 erlangs on direct link and remainder on link to transit.

However, if the traffic under investigation for the link is overflow type then dimension it to the heuristic overflow formula, rounding up the number of channels to the next 30 channel group size.

5.5 Summary of Chapter 5

The heuristic process using the three new algorithms can be described in figure 44.

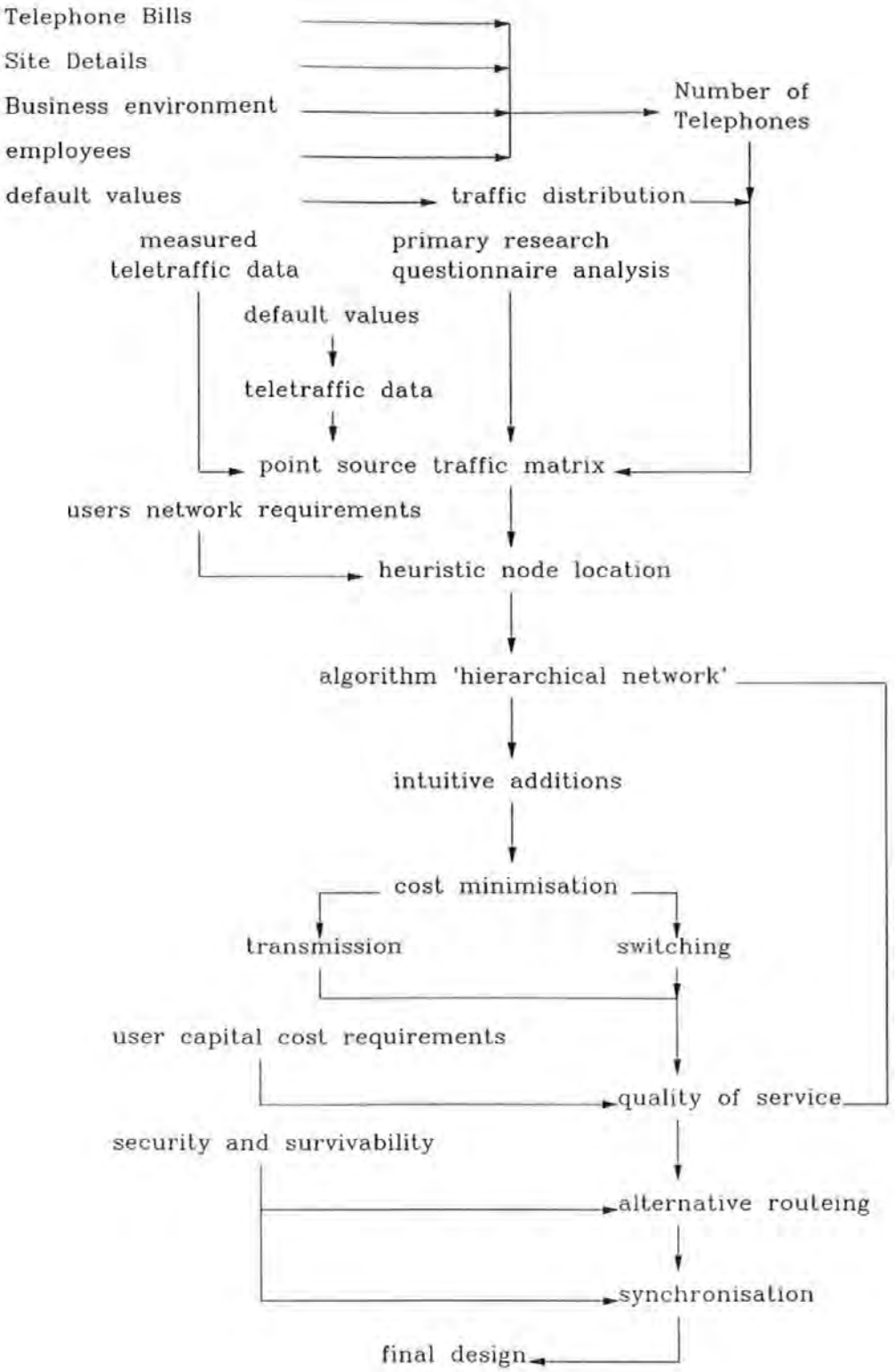


Figure 44 Heuristic Network Design Process

CHAPTER 6

A NEW APPROACH TO DEALING WITH UNCERTAINTY

6.1 Introduction

This chapter introduces the new concepts of 'Degrees of Integrity' and 'Uncertainty Windows'.

The uncertainty windows technique works by virtue of its ability to retain the composition of the uncertainty and allows for better solutions to be attained than with existing methods of dealing with uncertainty.

The characteristics of datum, fact, rule and inferred-datum windows are discussed, how these windows combine to produce an overall solution plane, and their eventual display to the user of the expert system is presented. This section of the chapter then concludes with a formal grouping of the windows into generic types.

The graphical concept of uncertainty windows modelling data, rules, facts and inferred-data is then illustrated, in the second section of the chapter, by way of a simple example.

6.2 Degrees of Integrity

Six degrees of integrity are considered necessary to represent adequately data, facts and rules in an expert system. They are:

i) uncertainty;

ii) reliability;

iii) relevance;

with inferred-data being represented by

iv) orientation;

v) intensity;

vi) profile.

These degrees of integrity can be represented graphically in a series of uncertainty windows which show them on either x or y axes as appropriate. This graphical representation technique allows the non-expert users of the expert system to have an insight and appreciation of the uncertainties that have helped to produce the overall system conclusion. It indicates which sources of uncertainty and which assumptions or rules are critical for further investigation to improve upon the confidence of the answer. The insight is qualitative in nature even if the expert system from which they derive is quantitative.

6.3 Uncertainty Windows

An expert system has two types of information contained within its database, one is generic and the other example-specific. The generic section of the database will hold all the rules, facts and subsequent inferred-data that are necessary for a problem solving in a particular domain, e.g. telecommunication network design. The example-specific section of the database will contain data on the particular problem to be

solved, e.g. a telecommunication network design for Bangkok. Generic data is entered into the system by experts when the expert system is compiled, whilst example-specific data is entered by the user of the expert system at the time when designs are required. The significance of the difference is that generic facts, rules, and inferred-data can have an associated set of complex uncertainty windows, i.e. they can have more than one degree of integrity. However, to maintain user-interface compatibility with existing expert systems, example-specific data has only a single degree of integrity, uncertainty. The result is that the two types of information need to be treated in different ways.

6.3.1 Data (Extrinsic to the Expert System)

Data is example-specific and is entered, by the user, as in conventional expert systems: for example, Datum($x \pm dx$), where datum is the information, x is its uncertainty, i.e. (1-Certainty), and dx is the uncertainty limits either side of the uncertainty value. This example is shown in figures 45 and 46.

Data has one degree of integrity:

uncertainty - based upon the users input, shown on the x-axis.

Data, as used by the Plymouth Expert System, maintains the normal 'dealing with uncertainty' characteristics as for existing expert systems; i.e. an uncertainty value of between 0 and 1 with a tolerance (or width).

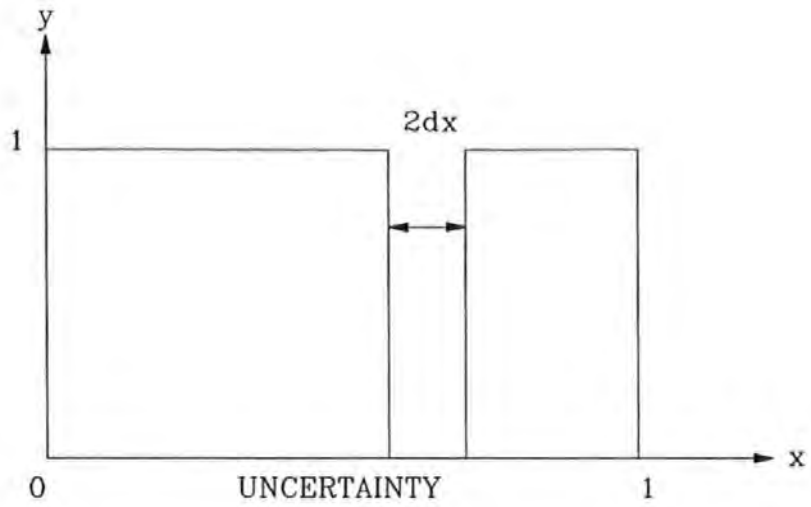


Figure 45 Datum Window

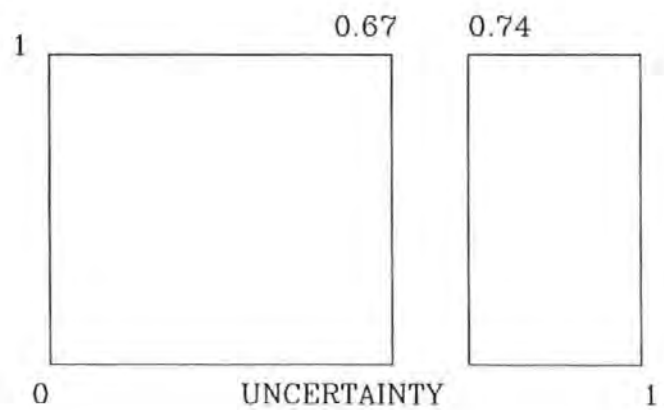


Figure 46 Datum Example

The model shows that data uncertainty windows exhibits a right orientation 'cut' for less certain data and a left orientation for more certain data.

6.3.2 Facts (Intrinsic to the Expert System)

Facts can be said to have three degrees of integrity:

uncertainty - similar to that of user data, shown on the x-axis;

reliability - dependent upon their source, shown on the x-axis;

relevance - dependent upon the application, shown on the y-axis.

These are shown graphically in figure 47.

Uncertainty is the first degree of integrity and acts with its uncertainties combined in some form, either added, multiplied, etc., with those associated with user data to contribute to the overall uncertainty profile.

Reliability is the second degree of integrity, it is derived from the quality of the source of the information, i.e. is he an expert, technician or layman? There exists a high correlation between reliability and confidence, indeed reliability can be redefined as 'confidence in source'. Thus the more confident one is in the source the less spread we have in the representation of confidence. Conversely, the less confident one is in the source, the greater the spread we must show in the confidence representation. This spread can be represented as a tolerance in the confidence. Thus a highly reliable source would have an associated cut in the uncertainty window of narrow tolerance

about the confidence value, whilst an unreliable source would be represented by a cut of wide tolerance.

Taking the example where our expert gave a confidence value of 0.3 (and hence an uncertainty value of 0.7) and we have a 90% confidence in the expert, then one has a spread in the confidence values of x±(1-reliability)/2, i.e. 0.7±5% or 0.7±0.035, as shown in figure 48. This is analogous to a mean and standard-deviation were the distribution to be Gaussian, but this is not necessarily the situation as the distribution could be of any form.

Taking the example further, the same confidence given by a novice in the field could result in only a 10% confidence in the source, resulting in an uncertainty window as shown in figure 49. The uncertainty window now has a cut on the x-axis with a value of 0.7±0.32; although 1.02 has to be truncated to unity, i.e. 0 and 1 form the boundaries between total certainty and total uncertainty.

Thus, as figure 48 and 49 show, the wider the cut, or tolerance, in the uncertainty window the less confidence we have in the source of the fact. Thus the more to the right the gap in the uncertainty window appears, the less confidence we have in the value of the fact itself.

Relevance is the third degree of integrity, it is represented by the 'depth of cut' into the uncertainty window on its y-axis. Totally irrelevant facts, no matter how certain the source or how confident we are in the source, have no part to play in the overall confidence in the inferred-data and will have a depth-cut of 100%. Highly relevant

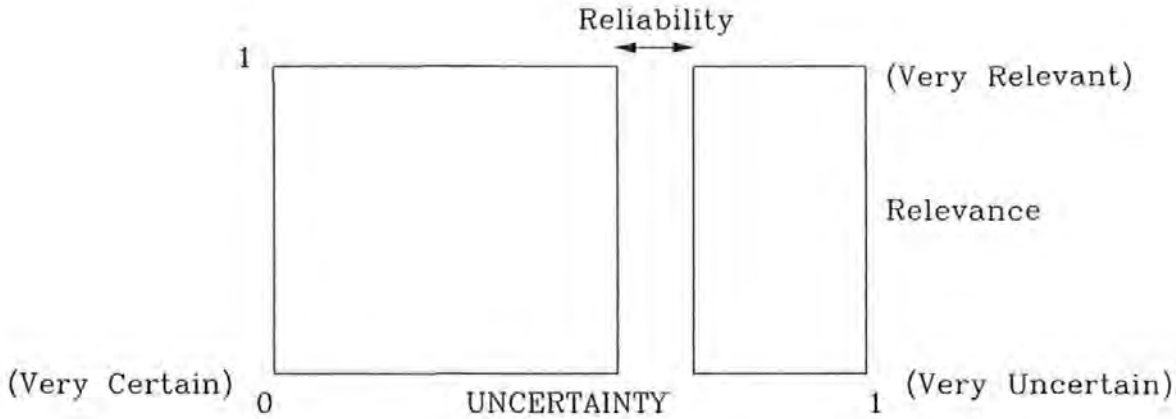


Figure 47 Fact Window

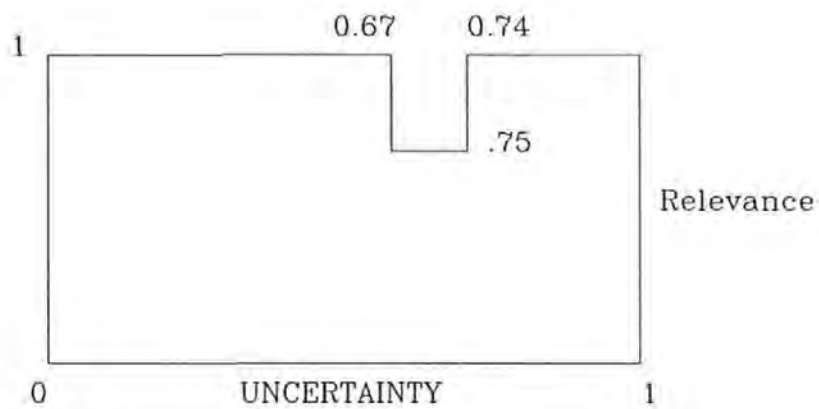


Figure 48 Fact Example 1

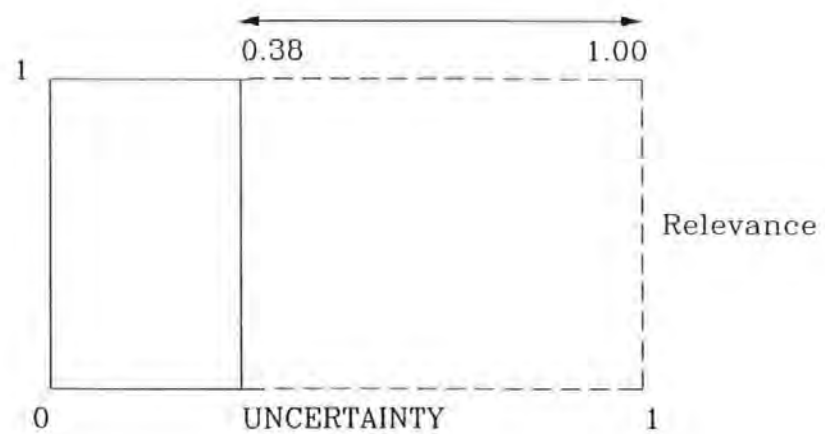


Figure 49 Fact Example 2

facts will have shallow depth-cut and contribute more to the profile of the resultant inferred-data.

6.3.3 Rules (Intrinsic to the Expert System)

A rough and ready heuristic rule may always seem to give the correct answer - but, without the rigour of mathematical proof, can we be sure that it applies in every case? This heuristic nature of the rule must reflect back into the overall inferred-data uncertainty profile, since the attributes of the rule are contained within the generic database and, as such, are invisible to the user.

Heuristic content is the first degree of integrity of a rule; the x-axis is used to represent heuristic content because of its direct analogy to confidence with regard to facts and data. Rules of a high heuristic content have right-oriented cut in the uncertainty window and rules that are recognised 'laws', left-oriented cuts.

Reliability is the second degree of integrity; it is a measure of the system designer's assessment of their confidence in the source of the rules, i.e. are they experts in the field? Reliability is directly analogous to its counterpart with facts: the x-axis or tolerance of the cut in the uncertainty window being used.

Relevance is the third degree of integrity and is directly analogous to its counterpart in the uncertainty window used to represent facts; the depth of cut into the uncertainty window represents the degree of integrity - relevance. In the limit an irrelevant rule will have a depth of cut of 100% and hence play no part in making the

profile of the inferred-data; highly relevant rules will have shallow depth-cut and contribute more to the profile of the resultant inferred-data, see figure 50 and 51.

In summary, rules have three degrees of integrity:

uncertainty - based upon the heuristic content, shown on the x-axis;

reliability - based upon the confidence of the source, shown on the x-axis;

relevance - based upon the application, shown on the y-axis.

6.3.4 Inferred-data (Intrinsic to the Expert System)

The process of generating the uncertainty profile of the inferred-data involves the superimposition of the uncertainty window of the rule, with those of its associated data and facts. The process is detailed in papers published by the author and these are reproduced in Appendix 3. The effect can be visualised as the shadows produced after light has passed through all the uncertainty windows concerned. The process is shown graphically in figure 52.

Degrees of integrity of inferred-data can no longer be thought of in terms of single identifiable contributors to the attribute.

Inferred-datum possesses three degrees of integrity; orientation or positional index, peakyness or intensity, and profile or width. These taken together give a confidence profile to the inferred-datum.

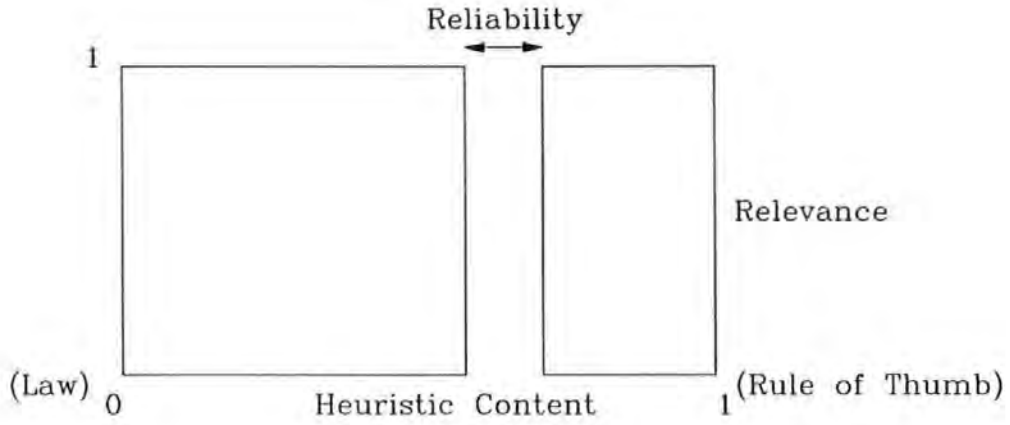


Figure 50 Rule Window

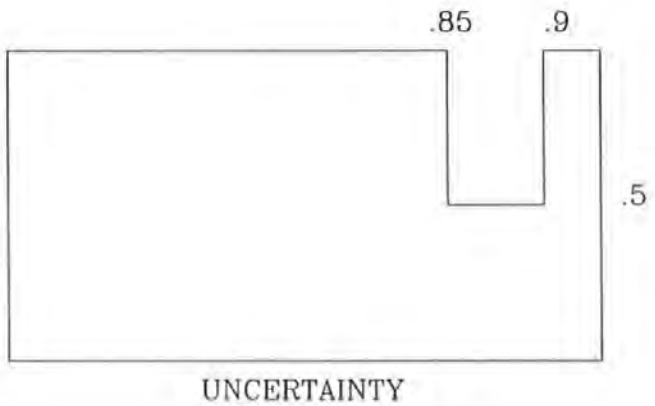


Figure 51 Rule Example
(e.g. Rule of Thumb - 50% relevant)

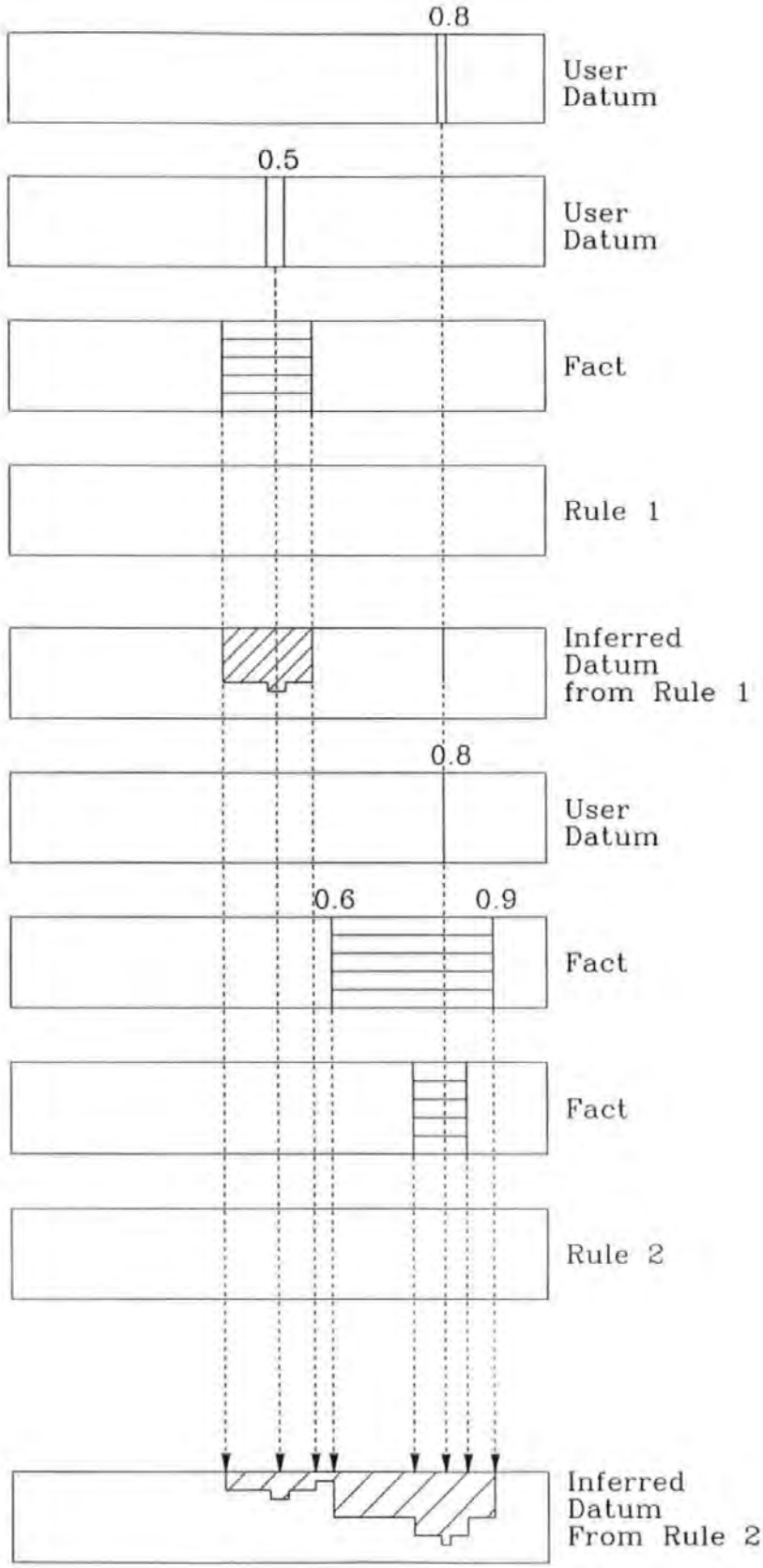


Figure 52 Inferred-Datum

Orientation of the resultant cut in the uncertainty window:

left orientation implies more certain inferred-data;

right orientation implies less certain inferred-data.

Intensity, resulting from the integration 'depth of cuts' in the uncertainty window:

deep-cuts imply non-relevant inferred-data;

shallow cuts imply relevant inferred-data.

Profile, resulting from the integration of 'cut-widths' in the uncertainty window:

narrow implies reliable inferred-data;

broad implies unreliable inferred-data.

Figures 53, 54 and 55 show resultant inferred-data profiles for differing input data and facts.

In summary, inferred-data have three degrees of integrity:

orientation, based upon uncertainty, shown on the x-axis;

intensity, based upon relevance, shown on the y-axis;

profile, based upon reliability, shown on the x-axis.

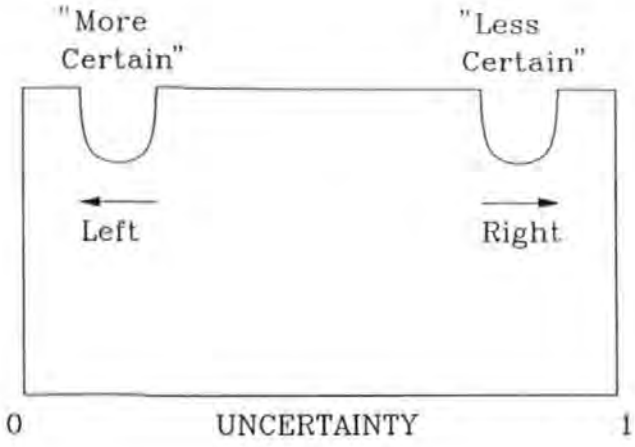


Figure 53 Inferred-Datum (Certainty)

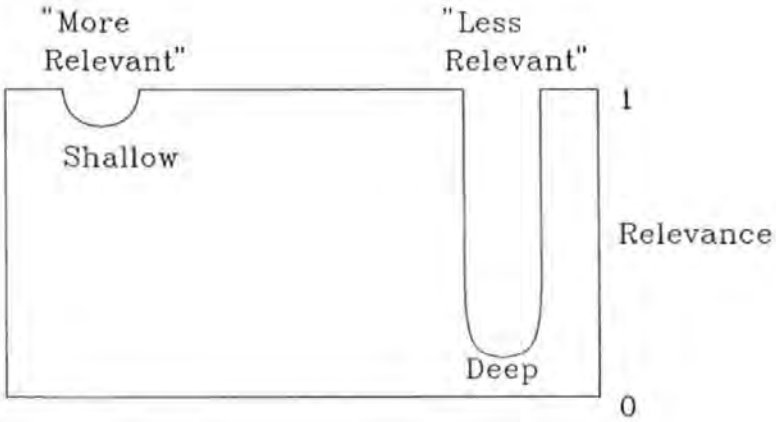


Figure 54 Inferred-Datum (Relevance)

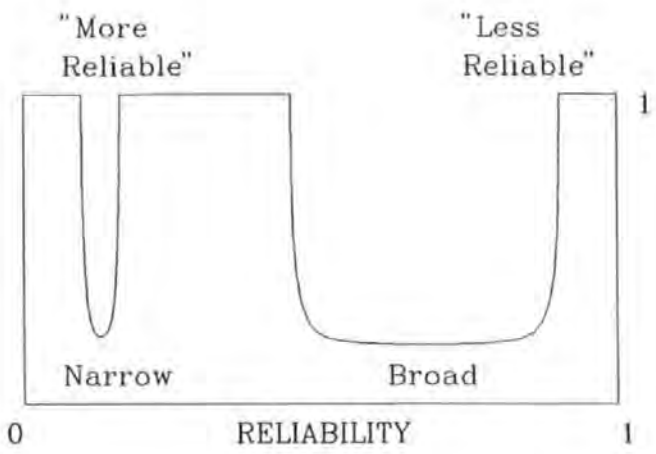


Figure 55 Inferred-Datum (Reliability)

6.3.5 The Solution Plane

The solution plane gives a visual indication of the quality of the solution produced by the expert system. The solution plane profile contains the detail of the uncertainty windows that were used in reaching to the conclusion. The construction of the solution plane is different from uncertainty windows in that rather than having windows with cuts representing relevance, reliability and uncertainty the solution plane contains details of each widow that contributed to the overall solution plane profile. This construction gives an additional dimension to an uncertainty window, z, which contains values representing the number of windows contributing to the total uncertainty at each point of the plane. (In other words and continuing the optical analogy, the solution plane contains the shadows representing the resultant confidence of the solution. Its form need not be the same as an uncertainty window, rather it can be considered to be the projection screen of the final inferred-datum. The solution plane thus represents the intensity of light after it has passed, or not as the case may be, through a series of uncertainty windows and cuts in the windows. The solution plane is a view of the density or intensity and position of light.)

To understand how the uncertainty windows unite to form a solution plane consider the following simple examples. The first example, figure 56, shows two uncertainty windows representing the uncertainty profile of rules A & B. The solution plane, a, holds the resultant elemental profile.

The second example, figure 57, shows how the solution plane is now considered to be input inferred-datum to two further rules C & D; a new solution plane, b, is produced, again the solution plane effectively stores the uncertainty windows from rules 1 to 4.

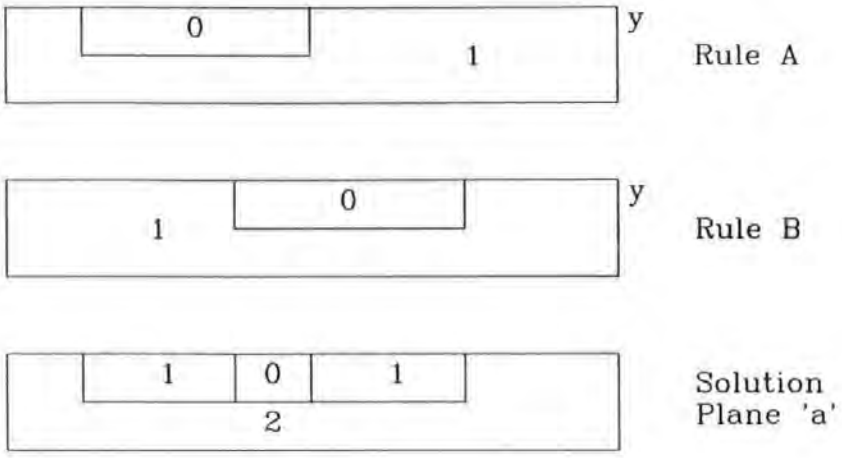


Figure 56 Combination Methodology (1)

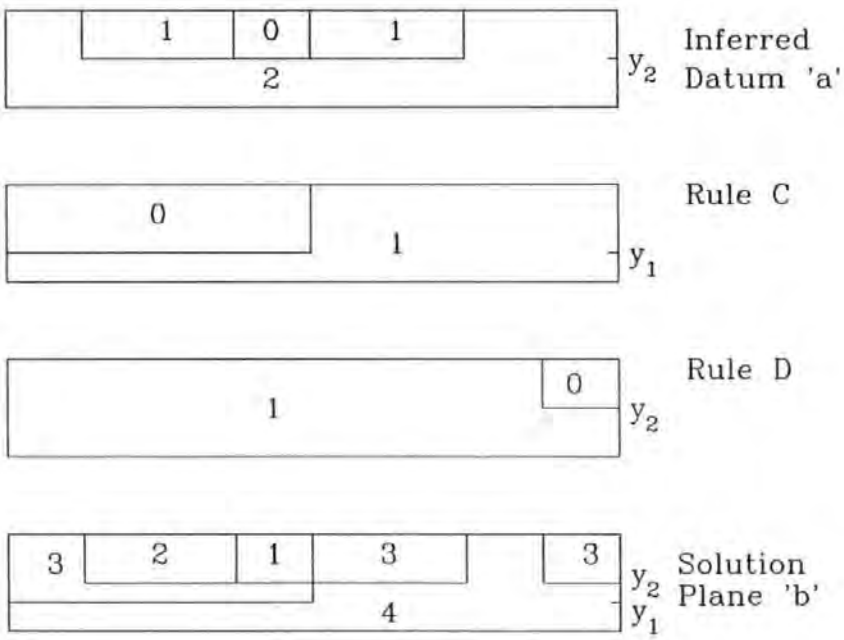


Figure 57 Combination Methodology (2)

These simple examples show that the solution plane can be presented to the user in two different forms: as an inferred-datum window and a resultant plane. The former can be considered an 'end-elevation' view and the latter a 'plan' view: figure 58 shows the two views. For analysis and display purposes, the solution plane is stored and updated in the form of memory elements, each memory element being represented by three pieces of information; its position as x and y co-ordinates and its elemental uncertainty, z. The number of elements necessary to represent a solution plane is a function of the number of rules in the expert system and the degree of granularity required.

Accurate interpretation of the solution plane is essential if the uncertainty windowing technique is to be successful. As an aid to interpretation it is valuable to present the solution plane, having three dimensions, in the form of a perspective graph. However, given the goal of achieving an appropriate balance between simple but uninformative, and the sophisticated but incomprehensible, one way of representing the elemental uncertainty, z, in a two dimensional plane, i.e. the computer screen, is for its representation by the allocation of a range of colours, from black for high to white for low and colours ranging from cold blue to warm red representing the elemental values between.

As a further aid to interpretation, masks can be employed each representing a profile that is progressively less acceptable. Figure 59 shows the general characteristics of a mask that could be employed.

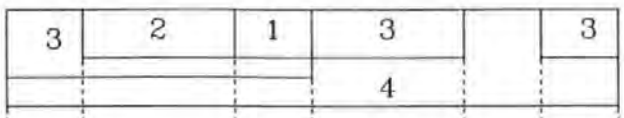
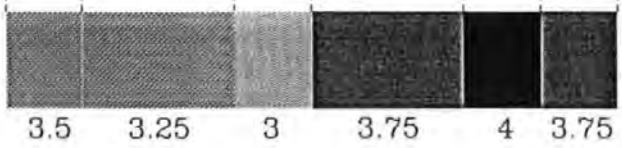


Figure 67a 'End-Elevation View of Solution Plane (Used for Inferred Datum to Subsequent Sub-Problem)



4 = Black
>0>4 = Grey (or colour)
0 = White

Figure 67b 'Plan' View of Solution Plane (Used for Presentation to User)

Figure 58 Two Views of a Solution Plane

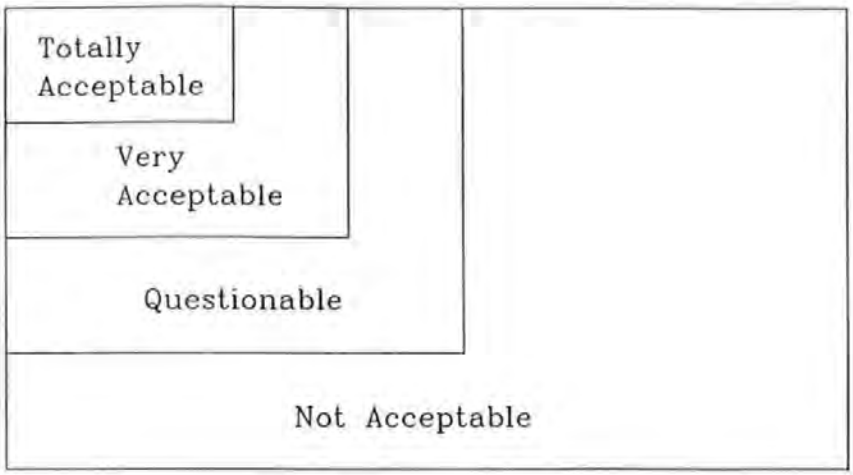


Figure 59 Inferred-Datum Mask (Outline)

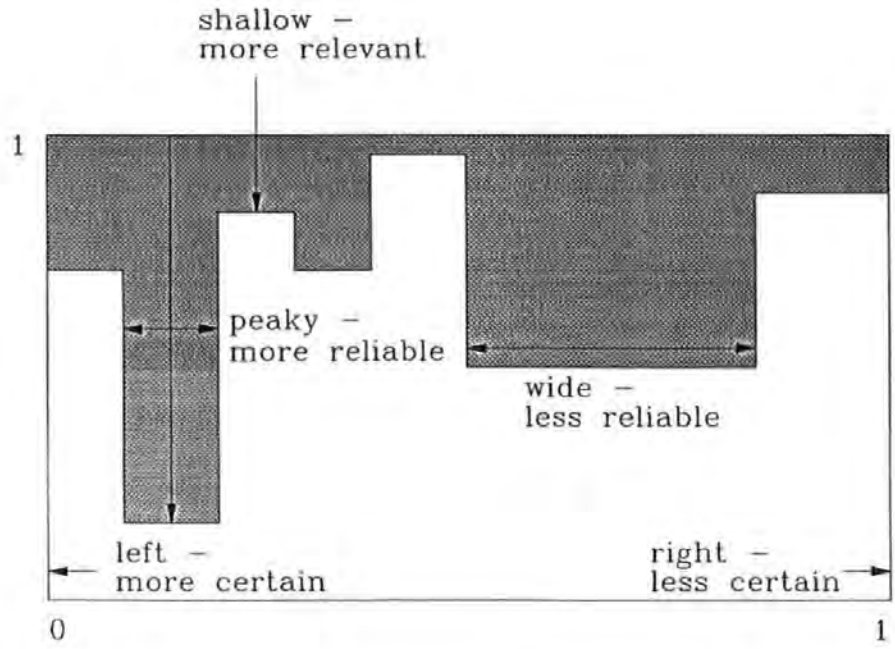


Figure 60 Uncertainty Window Characteristics

6.4 Uncertainty Window Combination Methodology

Figure 60 recapitulates the characteristics of an uncertainty window. To continue developing the combination methodology it is important that a resultant inferred-datum cannot be that is better than the intrinsic nature of the uncertainty windows used in its construction, i.e. a resultant inferred-datum window cannot be more certain or more reliable or more relevant than the rules, facts and data that have united to form the inferred-datum. For example, it is not possible to:

use only the overlaps of integrated cuts, as this would suggest that reliability has improved;

use resultant cuts which are shallower as this implies more relevant;

use resultant cuts whose widths are smaller as this implies more relevant;

move resultant cuts in windows to the left as this implies more certain.

The combination methodology must, however:

allow all uncertainty windows to contribute to the resultant overall profile;

not allow any one uncertainty window to dominate the resultant profile.

Using these guidelines coupled with the development of the example, in figure 57, by the addition of two further rules, Rule E (a totally perfect rule with no associated uncertainty) and Rule F (a totally imperfect rule of total uncertainty), it is possible to define the combination methodology.

Rule E has x and y values which are calculated as follows:

certainty = 1, thus uncertainty = 1-certainty, i.e. x=0;

reliability = 1, thus tolerance = x±(1-reliability)/2, i.e. x±0;

relevance = 1, i.e. y=1.

Thus the perfect rule uncertainty window is homogeneous, i.e. without cuts, and has a transmission value of 1, and whose uncertainty window profile is given in Figure 61. When superimposed on to the inferred-datum from rules A to D, whilst the resultant profile remains the same, the overall transmission value or density of each component that comprises the resultant solution plane is increased by one, i.e. it has the same profile but higher 'z' values.

Rule F shows the case of the totally unacceptable rule. Here its x and y values are calculated as follows:

certainty = 0, thus uncertainty = 1, i.e. x=1;

reliability = 0, thus tolerance = x±(1-reliability)/2, i.e. x±0.5;

relevance = 1, i.e. y=1.

In the case of reliability the tolerance has to be truncated at 1, as this is the right hand limit of the uncertainty window. Therefore the window width is from 0.5 to 1. This gives an uncertainty window profile as shown in figure 62; showing that the worst-case, rule uncertainty window, has a cut representing 50% of the total window width. Thus, the combination of a series of worst-case uncertainty windows does not result in the swamping, or submerging, of the profiles of the remanding uncertainty windows, but, rather attenuates the overall z values on the uncertainty windows' left hand side but not on the right hand side.

Figure 61 Totally Perfect Window

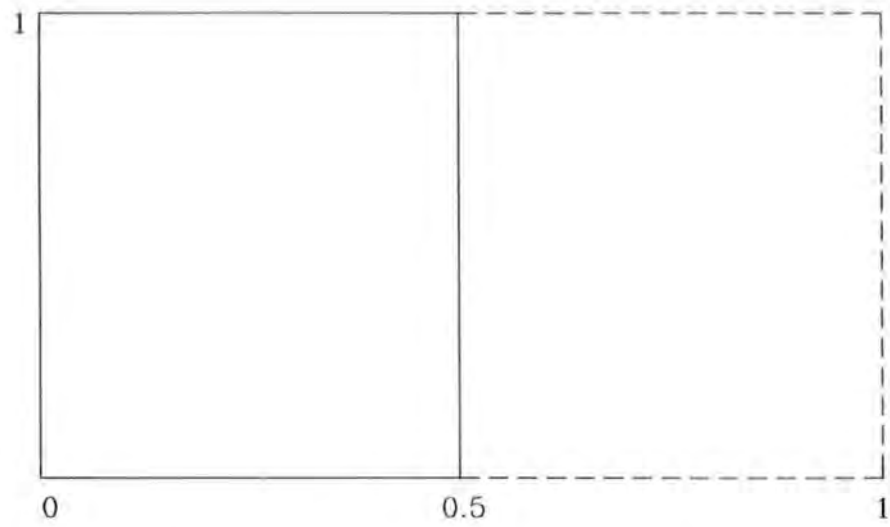


Figure 62 Totally Imperfect Window

Figure 63 shows the results of combining Rules A through E, whilst figure 64 shows the resultant uncertanity in its two forms: inferred-datum and solution plane.

6.5 The Uncertainty Window System in Operation

To demonstrate more fully the principle of the uncertainty windows in operation, a set of twenty heuristic-rules used to select the primary, secondary and tertiary master synchronisation node centres in a digital network, are worked through stage-by-stage. The illustration is based upon Case Study 1 - 'The Synchronisation Problem', detailed in Appendix 1.

The principle of the rule set operation is as follows:

the primary master should be located at the optimum 'centre' of a network;

the optimum 'centre' is when the sum of the individual attached routes are at a minimum synchronisation cost;

each route cost comprises link and node costs;

in each case a set of heuristics are used to give individual costs;

if there is more than one solution for the 'optimum centre'- then conflict resolution (traditional and uncertainty windows) is used.

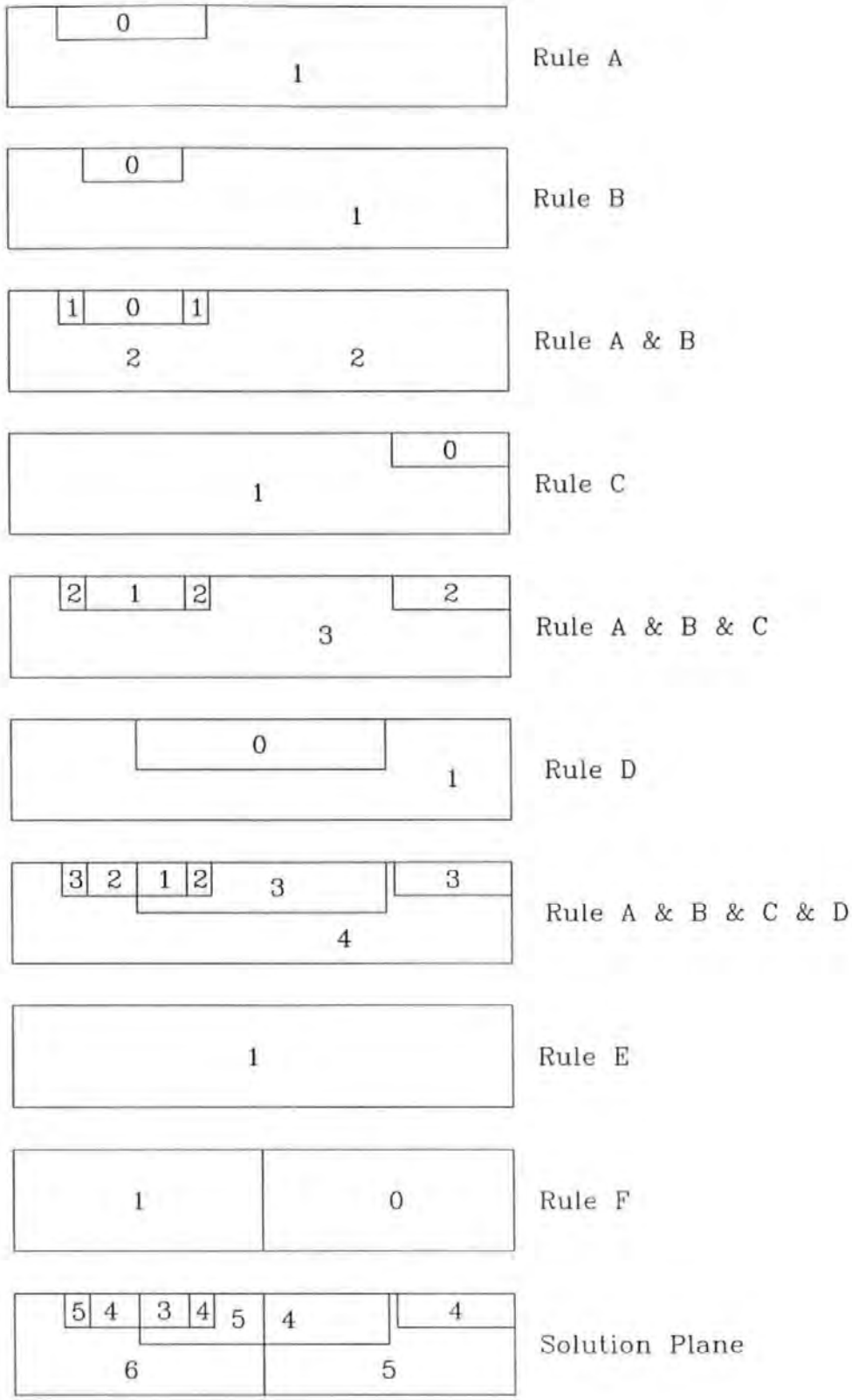


Figure 63 Combination Methodology Development (Example)

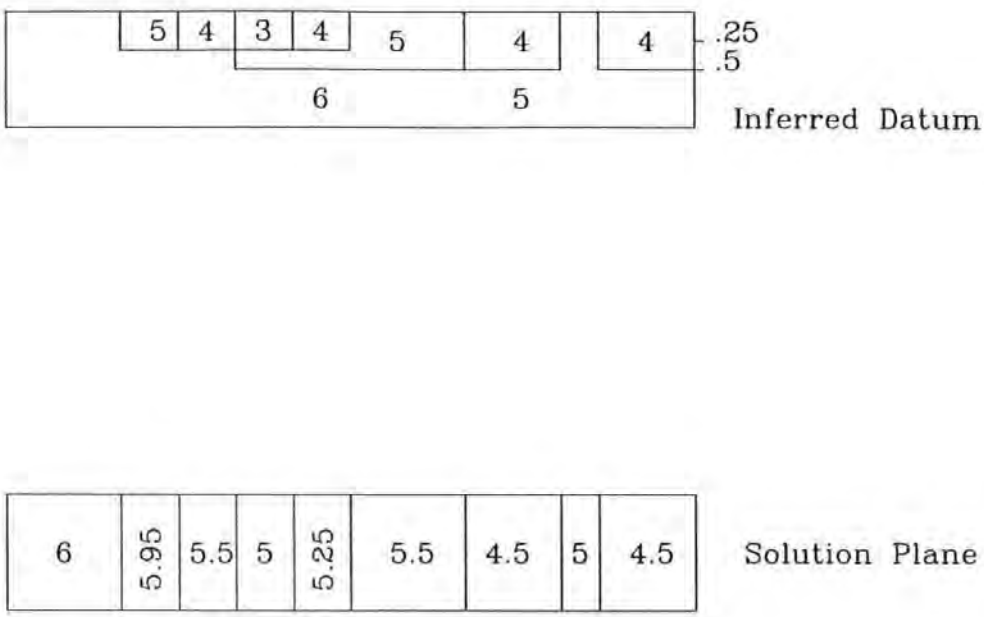


Figure 64 Solution Plane (Example)

The operation of the widows can be presented in a more formal context: the overall problem to be analysed by the expert system can be described as contained within a problem frame represented as [PROBLEM FRAME]. Within this frame are a number of sub-problem domains represented as {SUB-PROBLEM DOMAINS}. Within the sub-problem domains are the rules, facts, data and inferred-data relevant to that sub-problem and can be represented as {FACTS, RULES, DATA, INFERRED-DATA}. Within each fact, rule, datum and inferred-datum are its associated uncertainty window, represented as FACT(WINDOW).

Thus a simple problem would be described by:

[{Sub-Problem A(Fact A(Window), Rule A(Window), Datum A(Window); Inferred-Datum A(Window))}{Sub-Problem B(Fact B(Window), Rule B(Window) Datum B (Window), Inferred-Datum B(Window))}].

This problem representation is powerful as it is now possible to describe explicitly the interrelation of the sub-problems. Take, for example, the following hierarchical set of sub-problems:

[{sub-prob_a{sub-prob_b{sub-prob_d, sub-prob_e},sub-prob_c{sub-prob_f, sub-prob_g}}].

Here the output of sub-problem 'a' feeds to both sub-problems 'b' and 'c'; the output of sub-problem 'b' feeds to both the inputs of sub problems 'd' and 'e' and the outputs of sub-problem 'c' feeds to the inputs of sub-problems 'f' and 'g'.

When used for telecommunication network design, the Plymouth Expert System represents the problem, at its macro level as:

[Geographic Coverage{Penetration{Exchange{Direct, Graph, Synchronisation}, Traffic Matrix{Direct, Graph, Synchronisation}}, Building {Exchange{Direct, Graph, Synchronisation}, Traffic Matrix{Direct, Graph, Synchronisation}}, Customer{Exchange{Direct, Graph, Synchronisation}, Traffic Matrix{Direct, Graph, Synchronisation}}}];

where Geographic Coverage, etc. are the sub-problem domains of the Plymouth Expert System.

6.5.1 Windows in Action with Network Synchronisation

Pulse code modulation (PCM) transmission and all derived functions, such as digital switching, must be synchronised for correct operation. Individual PCM transmission links are kept in synchronisation, within themselves, by means of frame alignment. Here a demultiplexer extracts information (a frame alignment code pattern) from the incoming bit stream to derive an identical clock frequency at the receiver as at the transmitter.

In a digital network of switching nodes and links a complex arrangement is necessary to maintain complete synchronisation. The switching rates within a switching node may, typically, be governed by the frequency of a local quartz crystal oscillator of a Timing Unit (TU). This is a reliable frequency source but there will be minor differences between the natural frequencies of the TUs at the many switching nodes

in the network, since the natural frequency of the oscillators will drift from nominal with age and operating conditions. Thus the frequencies within a network of oscillators will change unless the network is locked (synchronised) to a single frequency source that has an acceptable margin of accuracy and reliability: the Primary Master Oscillator. The Primary Master must be backed-up with Secondary and Tertiary Masters for system reliability.

If the node clocks were not synchronised, the information rate of a signal received at a node would differ from the rate at which the node could process this information and possibly re-transmit it. This would result in information being lost (if the input rate was faster than the receiving node local clock) or repeated (if it was slower).

Within such a network there are various considerations that must be incorporated into the design. These include situations where the phase and frequency of the received clock vary in a cyclic fashion, called wander, where the physical path length of the link changes due to thermal effects: and where links or nodes fail, or are broken, and so are not available to pass on the synchronisation signals along the routes comprising a number of connected nodes and links.

The failure rate of a line system is length dependent, so the links contributing to synchronisation should be as short as possible. Links should also follow physical diverse routes where possible, so as to reduce the effect that a single fault could have on a designed network. Each link contributing to the synchronisation system requires a synchronisation unit (SU) to compare the phase of its local clock with that of the clocks received from the far end of the link. Each SU should be fail-safe to ensure that its malfunction has no adverse effects. A number of links participate in the

synchronisation control at each node and thus failure of an individual link or SU should not cause the clock at that node to lose synchronisation with the clocks at other nodes.

The size of actual networks and the complexity of traffic routeing usually makes necessary a hierarchical arrangement of nodes to give an economical solution.

Public Network Hierarchical Formalisation:

Level 4 = international digital node and primary master;

Level 3 = all transit nodes and the secondary master;

Level 2 = the tertiary master;

Level 1 = all local nodes.

If there is a network comprising only local nodes, then the resultant Private Network Hierarchical Formalisation is as following:

Level 4 = public digital node and primary master;

Level 3 = secondary master;

Level 2 = all PBXs interconnected in a hierarchical-net arrangement and the tertiary master;

Level 1 = all PBXs interconnected in a hierarchical-star arrangement.

6.5.2 Synchronisation Cost

In developing the network synchronisation rule-base, the idea of 'synchronisation cost' is introduced. Differing link and node technologies, line lengths, etc. are allocated a cost which ranges from zero, for the ideal, to infinite, for the unsuitable. The ideal solution to the design process is that with the minimum cost. Examples of the cost function contained within the Plymouth Expert System, with an indication of the cost principles and the constraining features, follow: whilst their derivation is detailed in Case Study 1, Appendix 1.

The cost of a link is a weighted function of the traffic capacity of the link, i.e., the larger the capacity, the greater the effect of failure. Thus the relative traffic cost is taken to be zero if more than thirty erlangs, otherwise 0.5.

The length of the link is proportional to the probability of its failure. Thus the relative distance cost is taken to be zero for distances up to 10 kms, 0.5 for distances between 10 and 1000 km and unity for distances over 1000 kms.

The cost of a link is a weighted function upon its ability to introduce wander into the synchronisation path. Thus the relative wander costs for differing link technologies are:

0	if <	1000 kms radio
0	if <	1000 kms buried cable
0	if <	100 kms aerial fibre
0	if <	10 kms aerial copper
0.5	if >=	1000 kms radio
0.5	if >=	1000 kms buried cable
0.5	if >=	100 kms aerial fibre
0.5	if >=	10 kms aerial copper
1	if >=	1500 kms buried cable
1	if >=	300 kms aerial fibre
1	if >=	60 kms aerial copper
1	if	satellite connection

The cost of a node is directly related to the number of digital links connected to it; as the failure of the node affects its neighbourhood nodes through the attached links. Thus the relative node cost is unity if the number of connected links is unity, 0.5 for between 1 and 10 links and zero for more than 10 links.

6.5.3 The Rule Base

The following rules are derived from the detailed explanations given in Appendix 1. However, the rules are shown in flow chart format in figure 65.

* Rule 1: The Primary Master should be located in the 'centre' of a network, where the 'centre' is a function of distance, traffic, link and node technologies.

* Rule 2: Centre of a network occurs when the cost of the maximum route budgets is at a minimum.

* Rule 3: The maximum route budget is the sum of the link and node costs taking the maximum cost between two nodes in a network.

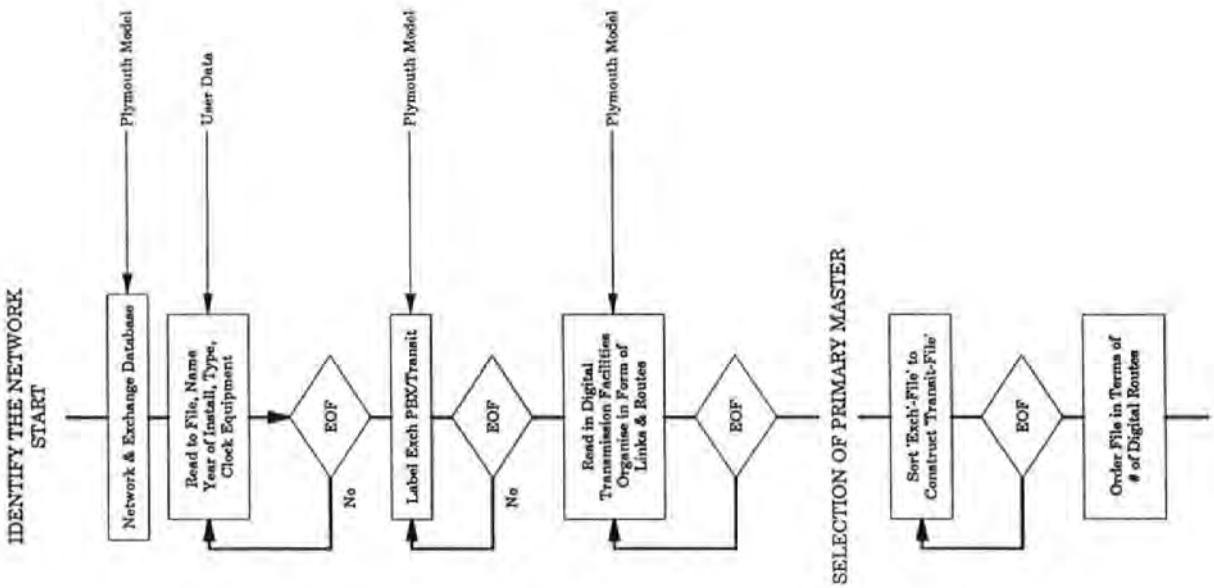
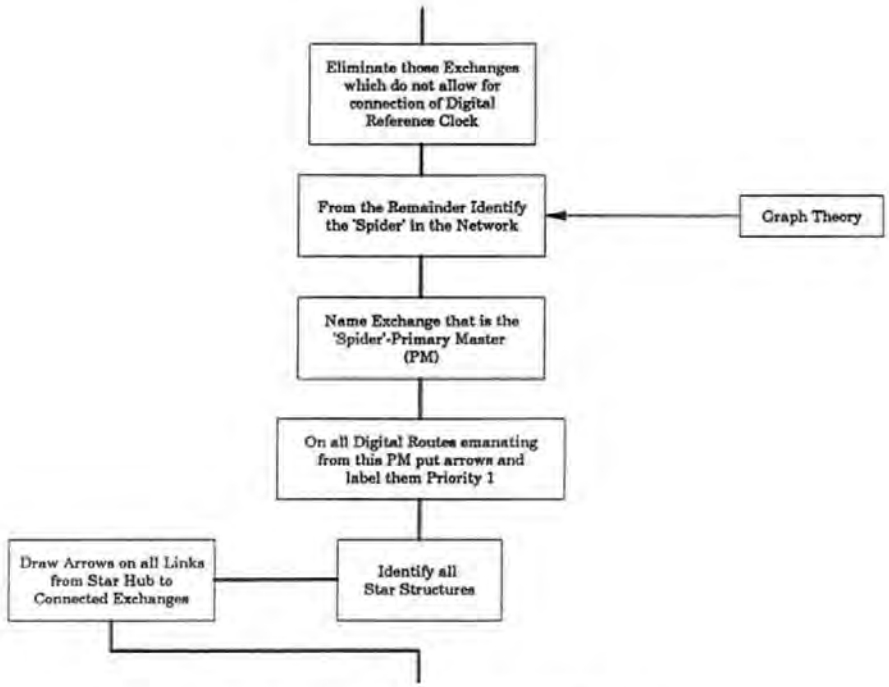
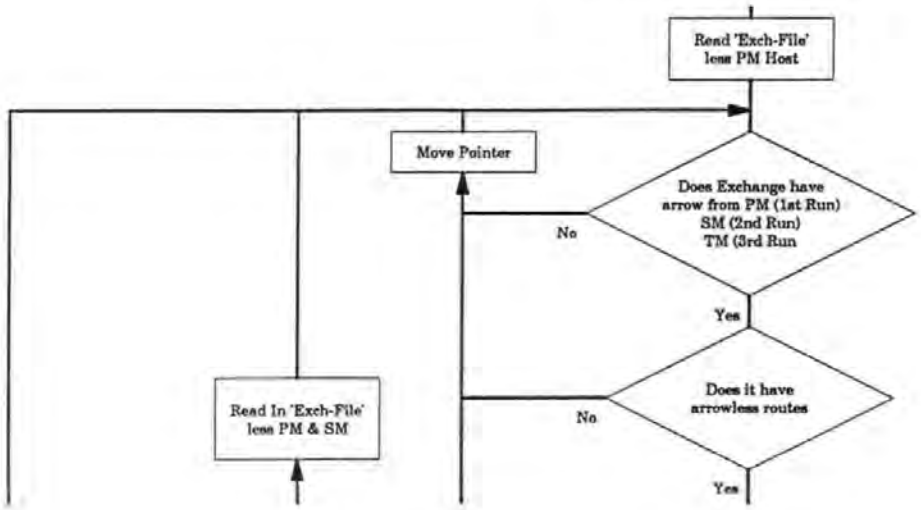
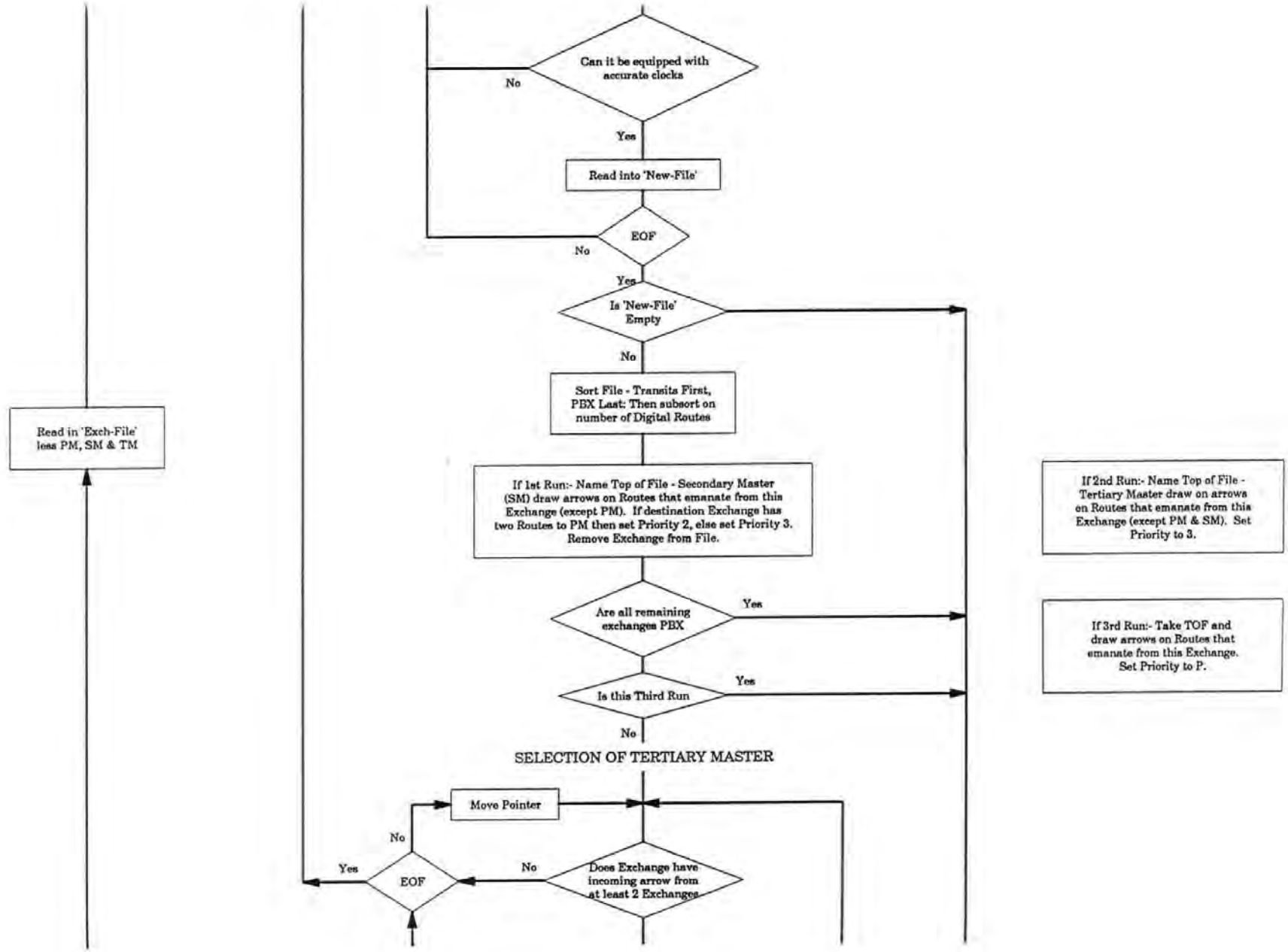


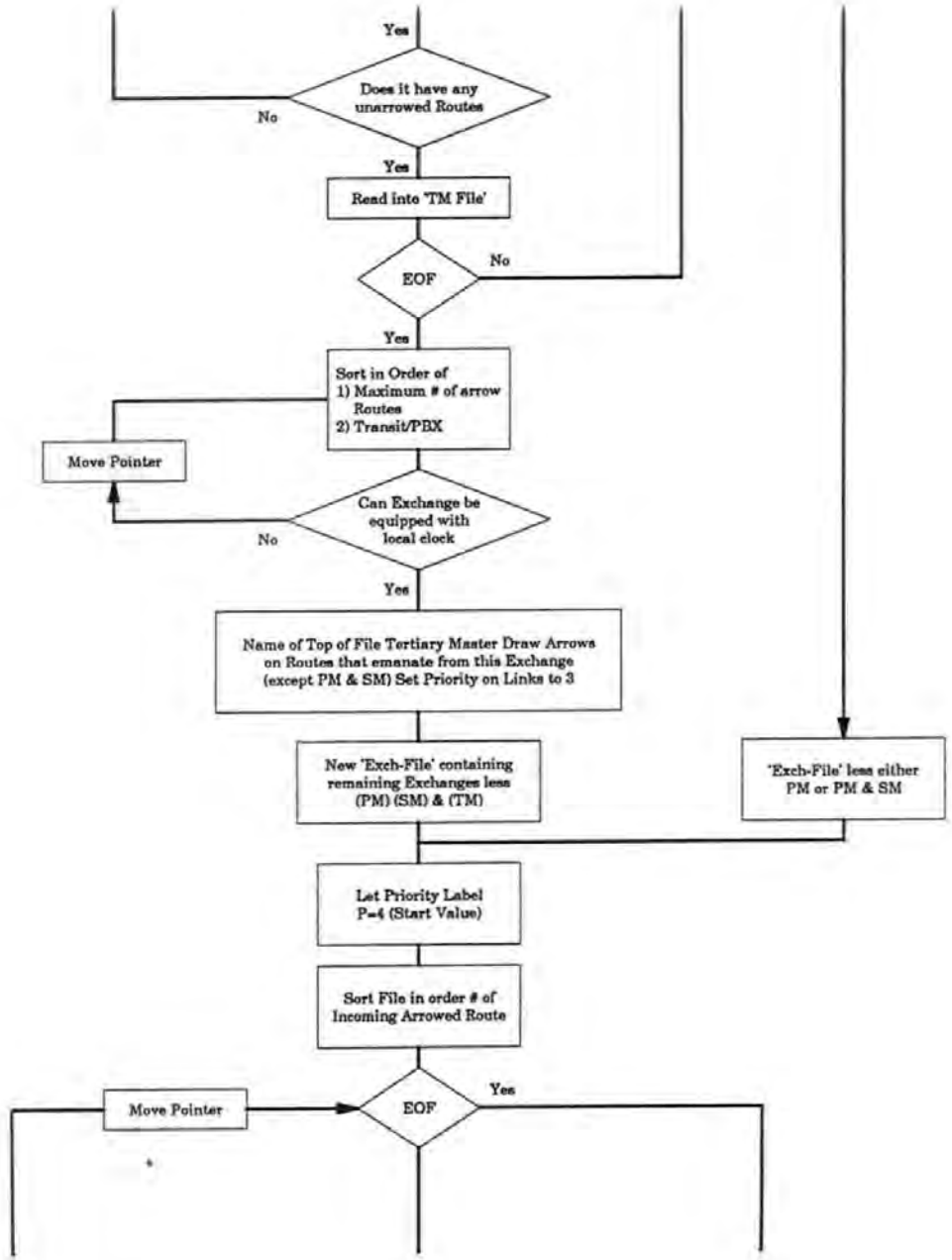
Figure 65 The Heuristic Algorithm for Synchronisation (Flow Chart)

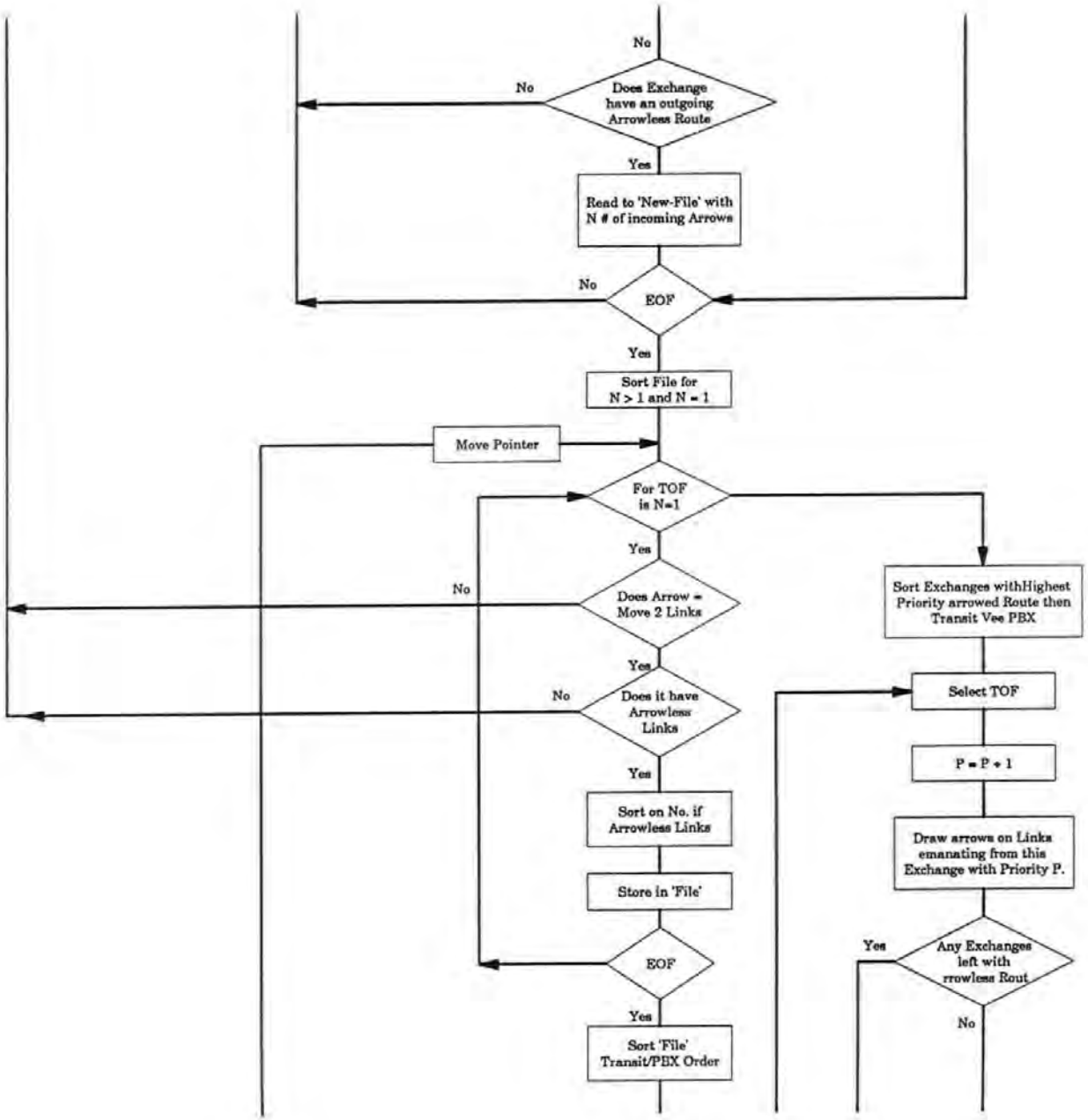


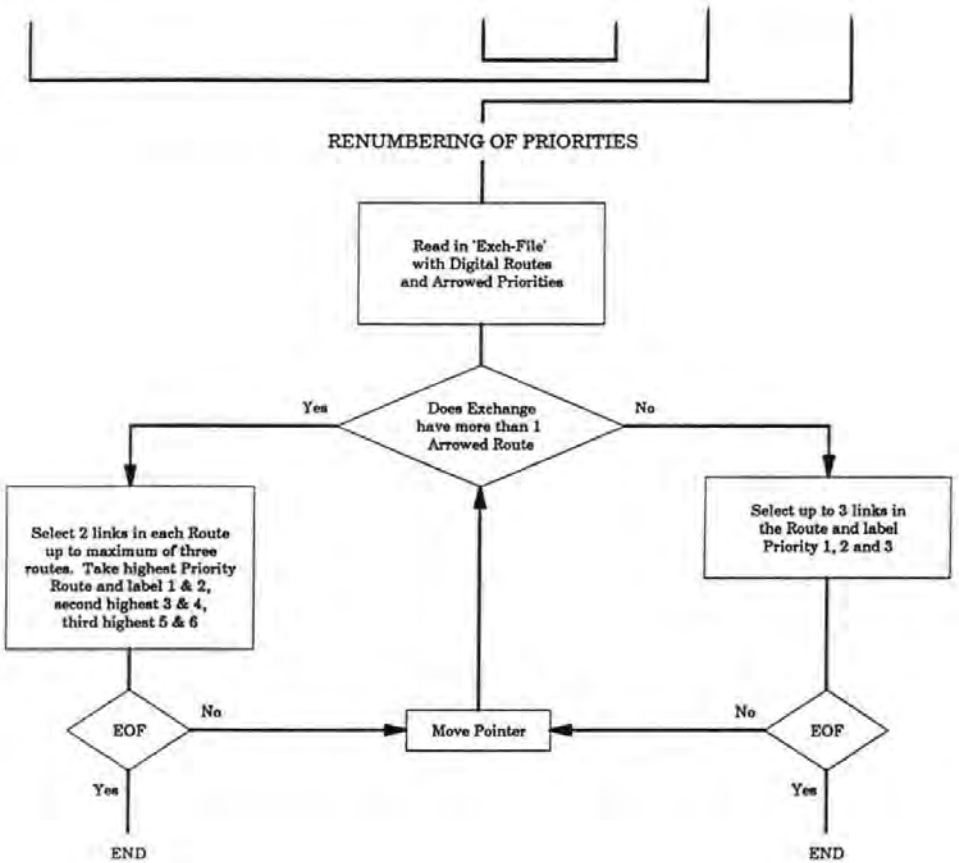
SELECTION OF SECONDARY MASTER




Read in 'Exch-File' less PM, SM & TM
Can it be equipped with accurate clocks
No
Yes
Read into 'New-File'
EOF
No
Yes
Is 'New-File' Empty
No
Sort File - Transits First, PBX Last: Then subsort on number of Digital Routes
If 1st Run:- Name Top of File - Secondary Master (SM) draw arrows on Routes that emanate from this Exchange (except PM). If destination Exchange has two Routes to PM then set Priority 2, else set Priority 3. Remove Exchange from File.
Are all remaining exchanges PBX
Yes
Is this Third Run
Yes
No
SELECTION OF TERTIARY MASTER
Move Pointer
No
Yes
EOF
No
Does Exchange have incoming arrow from at least 2 Exchanges
If 2nd Run:- Name Top of File - Tertiary Master draw on arrows on Routes that emanate from this Exchange (except PM & SM). Set Priority to 3.
If 3rd Run:- Take TOF and draw arrows on Routes that emanate from this Exchange. Set Priority to P.


Yes
Does it have any unarrowed Routes
No
Yes
Read into 'TM File'
EOF
No
Yes
Sort in Order of
1) Maximum # of arrow Routes
2) Transit/PBX
Move Pointer
Can Exchange be equipped with local clock
No
Yes
Name of Top of File Tertiary Master Draw Arrows on Routes that emanate from this Exchange (except PM & SM) Set Priority on Links to 3
New 'Exch-File' containing remaining Exchanges less (PM) (SM) & (TM)
'Exch-File' less either PM or PM & SM
Let Priority Label P=4 (Start Value)
Sort File in order # of Incoming Arrowed Route
Move Pointer
EOF
Yes


No
No
Does Exchange have an outgoing Arrowless Route
Yes
Read to 'New-File' with N # of incoming Arrows
No
EOF
Yes
Sort File for N > 1 and N = 1
Move Pointer
For TOF is N=1
Yes
No
Does Arrow = Move 2 Links
Yes
No
Does it have Arrowless Links
Yes
Sort on No. if Arrowless Links
Store in 'File'
EOF
Yes
Sort 'File' Transit/PBX Order
Sort Exchanges withHigheet Priority arrowed Route then Transit Vee PBX
Select TOF
P = P + 1
Draw arrows on Links emanating from this Exchange with Priority P.
Yes
Any Exchanges left with rrowless Rout
No


RENUMBERING OF PRIORITIES
Read in 'Exch-File' with Digital Routes and Arrowed Priorities
Does Exchange have more than 1 Arrowed Route
Yes
No
Select 2 links in each Route up to maximum of three routes. Take highest Priority Route and label 1 & 2, second highest 3 & 4, third highest 5 & 6
Select up to 3 links in the Route and label Priority 1, 2 and 3
EOF
No
Move Pointer
No
EOF
Yes
Yes
END
END

* Rule 4: The link budget comprises the sum of traffic capacity, wander resistance and distance costs.

* Rule 5: The cost of a node is a function of the number of digital routes.

Using the Rules 1-5 it is possible to identify one or more nodes that are network centres. If only one is found, it is necessarily the centre, otherwise the choice between them is determined by the greatest number of digital lines and account is taken of its level in the hierarchy.

* Rule 6: When two or more nodes have the same lowest maximum route cost the node with the greater number of digital routes is at the 'centre' of the network.

* Rule 7: When two or more nodes have the same lowest maximum route budget and the same number of digital routes, then the node that is higher in the hierarchy is the 'centre' of the network.

* Rule 8: When Rule 7 has failed to select the 'centre', the sum of the distances of every connected link to each node is calculated. The network 'centre' is now the node having the minimum distance cost.

* Rule 9: When Rule 8 fails to select a 'centre', the technology of the links is considered and the 'centre' taken as the node having the minimum wander cost on the links.

* Rule 10: If, in Rule 9 there are a mixture of technologies in a single route, the worst case is taken as representing the whole distance.

By this stage, the Primary Master location will have been found. It is then necessary to discover the Secondary and Tertiary Masters by repeating the above process. However, at this stage it is not necessary to find the optimum 'centre' but rather the next best two node locations that are closest to the Primary.

* Rule 11: Identify the best two neighbour nodes to the primary master node and search each, re-using Rules 1-10 replacing the master node with a cost value by zero.

The interconnection of nodes must be made to a set of rules based on the hierarchy. In addition, a number of resilience requirements must be satisfied.

* Rule 12: A node can only give a synchronisation to another node at the same level or lower level in the hierarchy.

It is also necessary to prevent looping thus the direction of control is important. Under normal operating conditions the Primary must have precedence over the Secondary and it over the Tertiary.

* Rule 13: Control between Masters will be unilateral from the Primary to the Secondary to the Tertiary.

* Rule 14: Control between nodes at differing levels in the hierarchy is unilateral from the higher to the lower.

* Rule 15: Control between nodes at the same level is bilateral.

The requirement for resilience dictates the following link requirements.

* Rule 16: Every level 3 node should be connected to a level 4 node by no less than two synchronisation links in tandem and the precedence allocated to the links is greater to those links from the higher level.

* Rule 17: If there are more than one link from the same level terminating on the node, a higher precedence is given to the link that has the lowest number of subsequent links in tandem to the primary master.

* Rule 18: Level 2 nodes should have a minimum of two synchronisation links with precedence given to the link from the higher level.

* Rule 19: Level 1 nodes should receive their synchronisation from a node at level 2. If this is not possible then a minimum of two links from the same level is preferred.

* Rule 20: Where no suitable primary rate link is available, a 64 kbit/sec link to the highest level with which the node has a traffic community of interest is required.

6.5.4 The Rules in Operation

To simplify the example, link costs are set to unity and node costs to zero. See figure 66 - and the table below:

Start Node	By Rule 3 cost	By Rule 3 route	By Rule 6 digital links	By Rule 7 hierarchy
1	5	1-2-3-5-4-6	2	1
2	5	2-1-3-5-4-6	3	2
3	5	3-1-2-4-5-7	3	2
4	5	4-2-1-3-5-7	3	3
5	5	5-3-1-2-4-6	3	3
6	6	6-4-2-1-3-5-7	1	1
7	6	7-5-3-1-2-4-6	1	4

Notice that nodes 1 to 5 have the same minimum cost (5) and that conflict resolution will be necessary to select the network centre. Rule 6 is used to reduce the number of 'acceptable' centres from five to four, i.e. nodes 2 to 5 inclusive. By Rule 7 nodes 4 and 5 may be eliminated, reducing the number of acceptable centres to nodes 2 and 3 only. Rule 8 may now be used by considering the distances involved. The example has now to be enhanced with uncertain data as shown in the table below.

User Defined Distance	User's Confidence	Involved Node
1-2⇒1	0.9	2
1-3⇒1	0.9	3
2-3⇒1	0.9	2 and 3
2-4⇒2	0.8	2
3-5⇒2	0.7	3
4-5⇒1	0.7	
4-6⇒3	0.5	
5-7⇒3	0.5	

A traditional method of dealing-with-uncertainty would give:

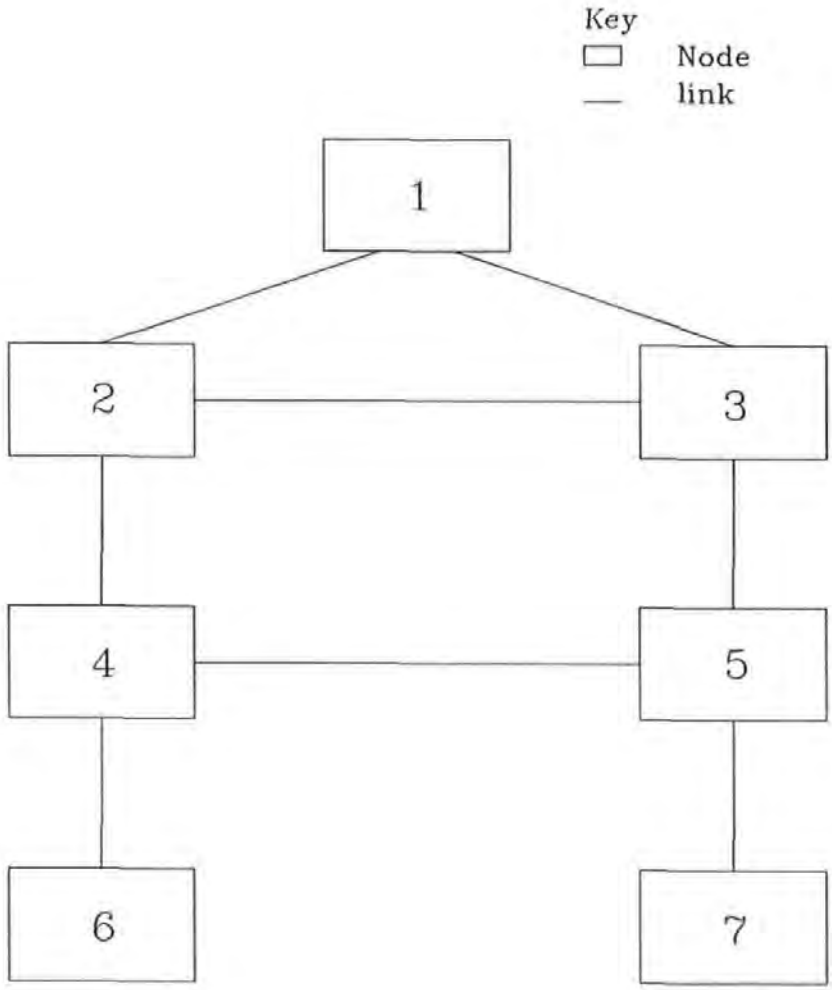


Figure 66 Example Network (Synchronisation)

node 2 - distance sum 4, confidence 0.648

node 3 - distance sum 4, confidence 0.567

Unlike the traditional method, the uncertainty windows method, shown in figure 67, shows that the original confidences attached to the data are not now lost but reflected in the confidence profile of the inferred-datum, with the orientation and intensity degrees of integrity of the inferred-datum clearly shown. The actual inferred-datum values in this case are determined from the individual resultant values, (1+1+0)/3=0.67 and (0+0+1)/3=0.33. However, we are still unable to select the correct centre from Rule 8 since both distances equal 4 and the confidence profile is approximately the same in both cases. Rules 9 and 10 are therefore used in consideration of the technology, as shown in the following table.

Links	User Defined Technology	User's Confidence	Involved node	By Rule 9 Cost
1-2	900 kms. radio	0.2	2	0
1-3	900 kms. aerial fibre	0.95	3	1
2-3	1000 kms. aerial fibre	0.8	2 and 3	1
2-4	2000 kms. radio	0.5	2	0.5
3-5	2000 kms. fibre	0.8	3	1

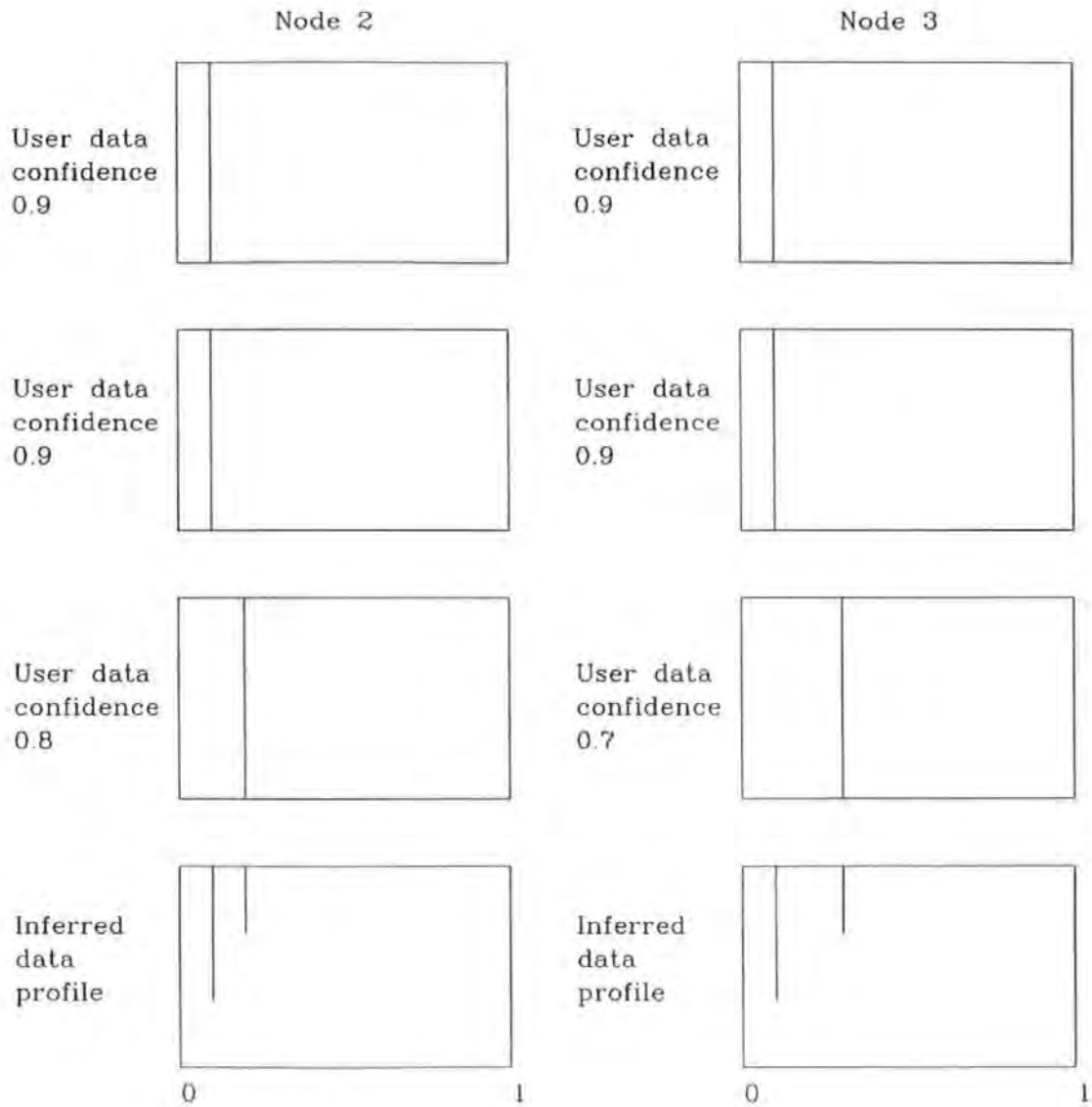
A traditional method of dealing-with-uncertainty would give:

node 2 - sum of costs 1.5, confidence 0.080,

node 3 - sum of costs 3.0, confidence 0.608,

node 2 being the network centre with its associated lowest cost.

But, how does one deal with the associated low certainty of 0.080?



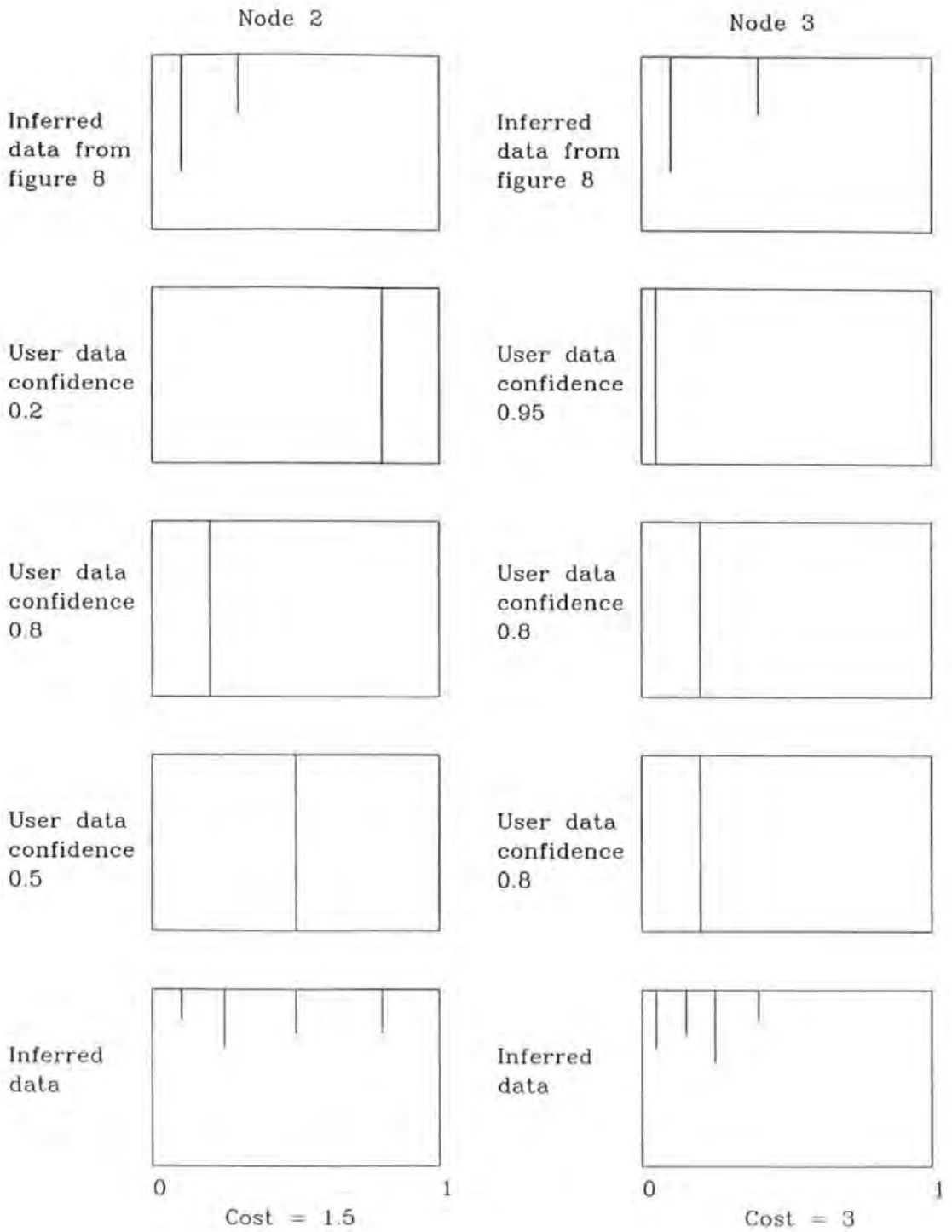
In each diagram, horizontal scale: uncertainty; vertical scale: relevance

Figure 67 Uncertainty Window Selection (1)

The uncertainty window method, shown in figure 68, indicates that much of the data supporting the choice of node 2 has confidence values of 0.5 and greater, which suggests that the calculated confidence value of 0.08 is misplaced. There is enough high confidence data in the conclusion for node 2 to support its claim to be the network centre on grounds of lowest cost.

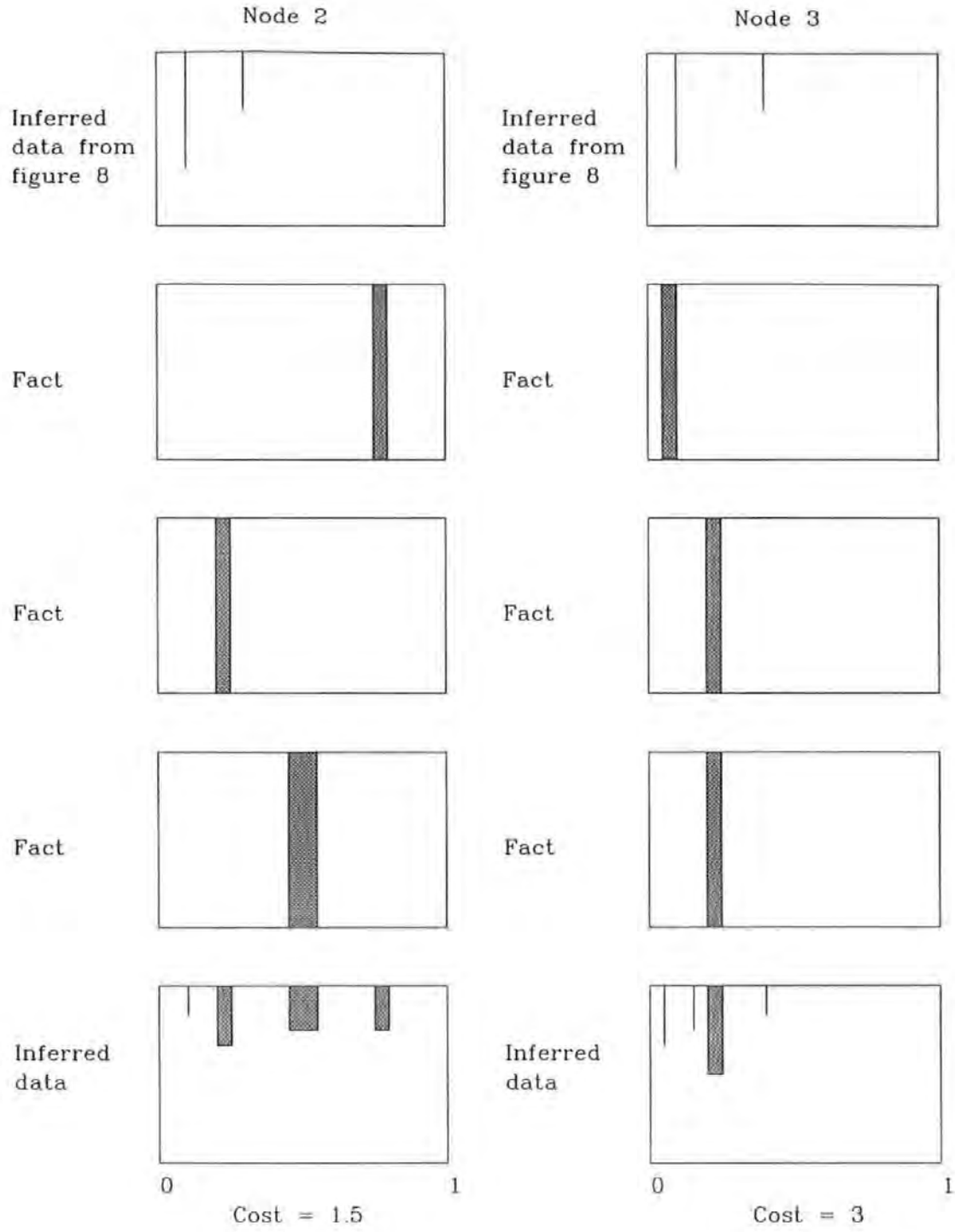
Figure 69 shows the result of including the reliability degrees of integrity. Here the spread of integrity of facts is passed into the inferred-datum with its values determined as in the previous figures. The resultant inferred-datum with its associated three degrees of integrity for orientation, intensity and profile supports the selection of node 2 as both relevant and reliable facts give a left-oriented, shallow and narrow profile. This example is rather simplistic and in practice there is normally overlapping and dispersion over part of the confidence range. Even so, there is often a concentration of confidence, or it can be obtained by producing more appropriate reliable data and facts to help with an indication of the best choice for the specific application.

Figure 70 shows the solution planes of both nodes 2 and 3.



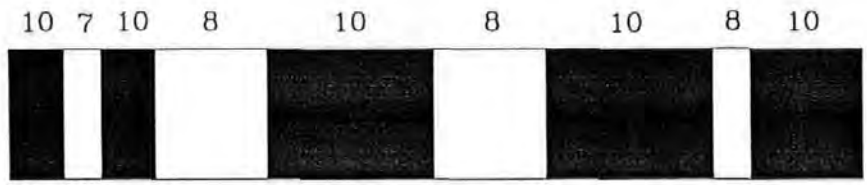
In each diagram, horizontal scale: uncertainty; vertical scale: relevance

Figure 68 Uncertainty Window Selection (2)

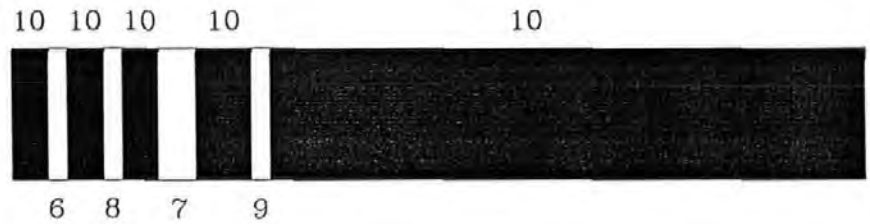


In each diagram, horizontal scale: uncertainty; vertical scale: relevance

Figure 69 Uncertainty Window Selection (3)



Node 2 Cost = 1.5



Node 3 Cost = 3

Key

10 = Black

0 = White

>0<10 = Grey

Figure 70 Solution Planes for Nodes 2 & 3

CHAPTER 7

SYSTEM IMPLEMENTATION & RESULTS

7.1 The Development Environment

The combination of Windows, FoxPro, C, SD-Prolog and Graphic-Server software provided a development environment which proved suitable for the development of an expert system appropriate for telecommunication network design. This was because the problem has a range of demands that were best suited to different computer languages.

The first demand was for an interface with which a user felt comfortable and found easy to use; Windows provided that user friendly interface to the system.

The second demand was for a database necessary to hold the large amounts of information required to describe the telecommunication network. FoxPro provided this function; it is a Windows based database with interfaces to other programming languages.

The third demand was for the support of the purely computational requirement of the network design; for example the calculation of link sizes, estimation of demand from GDP, etc. The C language provided the best option in terms of compatibility with the operating system and the database.

The fourth demand was for a language that supported heuristic search and design: SD-Prolog was a natural choice with its in built-in rule based format and search mechanisms.

The fifth demand was for a graphics package which would facilitate the presentation of uncertainty windows to the user of the system. Graphic-Server, a graphics development tool for Windows, provided this functionality.

The Plymouth Expert System requires that all languages are either to be available at run time or held as separate executable files. However, when used for commercial applications, the system would need to be compiled into a single executable file that would run under Windows.

The runtime of the programs that comprise the Plymouth Expert System is short in comparison with other computerised network modelling tools. For example, when run on a 386 personal computer with 8 Mega-bytes of memory, the design of a network for the city of Bangkok takes just over thirty minutes to complete; starting from the raw data representing the market for telecommunication services, in a mesh of 250*250 metre blocks, to a final design which includes the node locations and interconnection network.

7.2 The Expert System

In designing the Plymouth Expert System, it was a primary objective that it would support both the network design requirements and provide a platform for the analysis of uncertainty calculus. Those parts of the network design process that contributes

most uncertainty, i.e. traffic matrix generation and synchronisation planning, are detailed as case studies appended to this thesis. Communicating the results of the modelling process was also a critical part of Plymouth Expert System: a graphical display of uncertainty windows was developed that allowed the communication of insight, including an appreciation of the overall uncertainty about the conclusion; an understanding of which sources of uncertainty and which rules, fact and input data are critical to those conclusions and an understanding of the extent to which plausible alternative assumptions can change or improve the confidence in the conclusions obtained. The uncertainty windows are non-monatomic in their nature, a very desirable feature as it allows conflicting evidence to be assimilated.

The Plymouth Expert System is controlled by using domains built into the system. Domains limit the area where rules apply. This is necessary because once the knowledge base has more than about 100 rules, it becomes very difficult to grasp how it is operating and the response can become rather slow. This feature is implemented by calling up new knowledge bases as necessary. However, if the domain is unable to reach a satisfactory conclusion the rule set of other domains are checked so that rules can be used outside their normal mode if this helps to draw conclusions; use of these rules is noted in the relevance attribute, i.e. y-axis, of the rules' uncertainty window.

7.3 The Plymouth Expert System in Use

The best way to describe the Plymouth Expert System is by way of an example. To explain the systems operation it is said to operate in three phases: phase 1 takes the raw input data and changes it in to a form that is in a traditionally recognisable form, i.e. number of telephones, calling rates, etc.; phase 2 coverts the refined data into a

traffic matrix; phase 3 selects the node locations and the interconnecting links. Figure 71 shows the input screen of the Plymouth Expert System.

7.3.1 Phase 1 of the Plymouth Expert System Operation

Three problem domains dominate phase 1 of the Plymouth Expert System. They are called Penetration, Building and Customer.

The Penetration domain requires the area under examination to be divided into 250*250 metre blocks, see Figure 72. The telephone densities are then calculated based upon the primary, secondary and tertiary land utilisation within the square, its relevant penetration factors and location, i.e. city core, suburban and city fringe. The estimation of utilisation percentages will have associated with it a significant uncertainty, as rough estimates are calculated based upon visual interpretation of the percentages.

Traffic calling rates are then calculated based upon the profile of individually named customers in known locations. Uncertainties become compounded by the multiplication of penetration factors and stereotypical profiles based upon averages. Indeed, these averages can be calculated from historical data produced from a number of previous network designs. The penetration factors and their associated uncertainty windows are pre-programmed into the expert system; they have two attributes, heuristic content and reliability. Relevance is not programmed at this stage and is introduced, at run-time, when the rule is used outside the sub-problem domain for which it was intended. The resultant inferred-data output has associated with it its set of uncertainty windows.

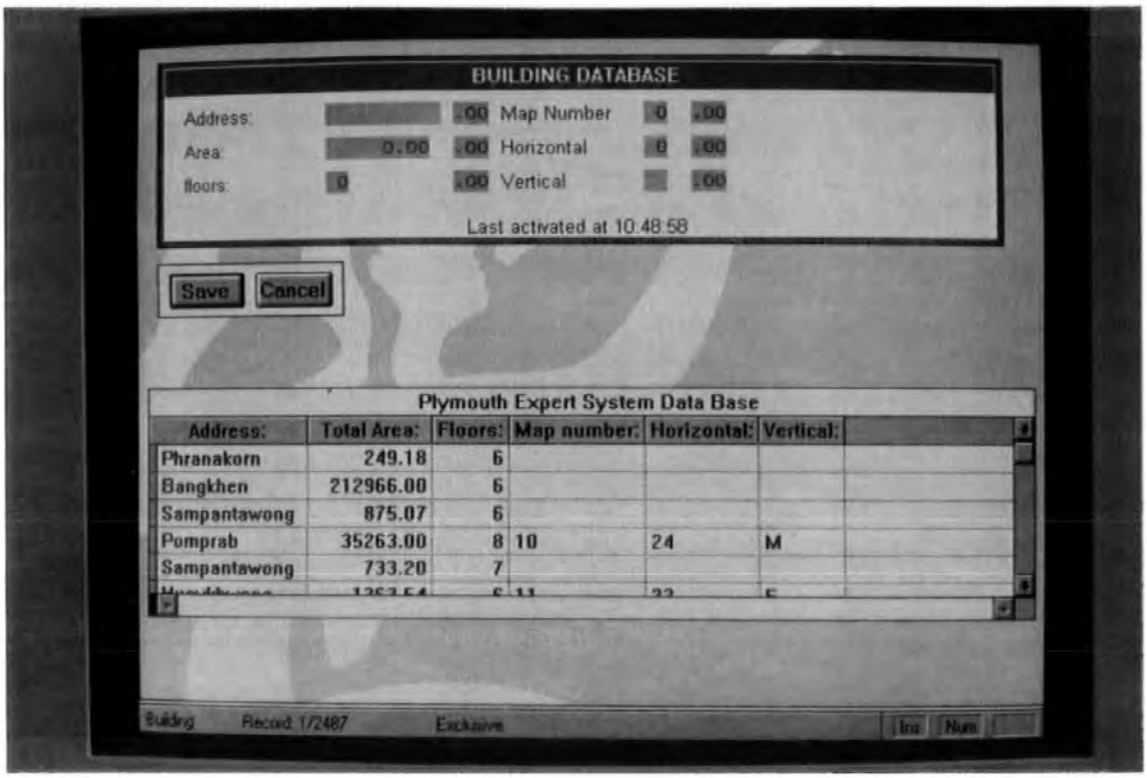
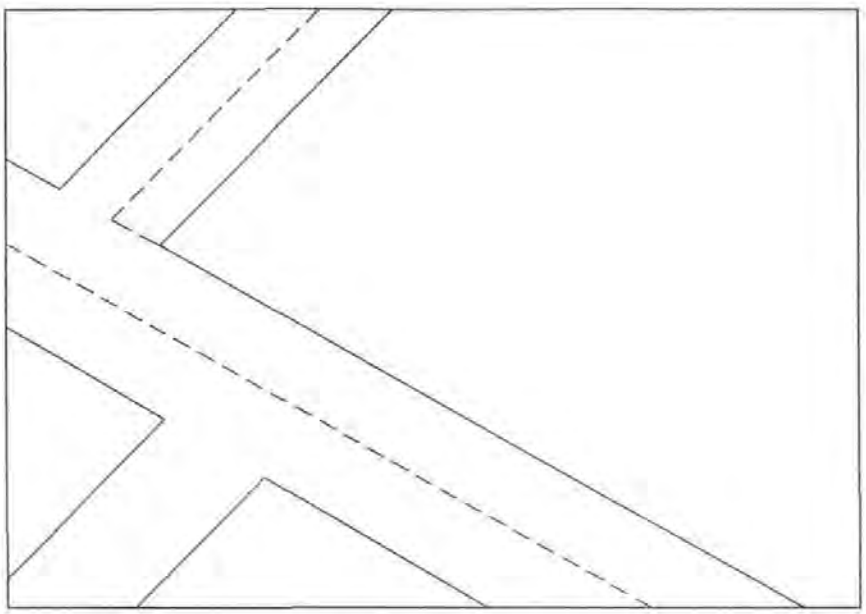


Figure 71 Plymouth Expert System Input Screen



% Coverage of Low Residential
High Residential
Commercial
Agricultural
Factory
Government

e.g.	Primary	Secondary	Tertiary
	High Residential 60%	Low Residential 15%	Factory 5%

Use general penetration factors to give

(1) No. households per block [– core, – suburban, – fringe]

(2) No. telephones per household

(3) Demand = (1) x (2)

(4) Traffic = (3) x general calling rates

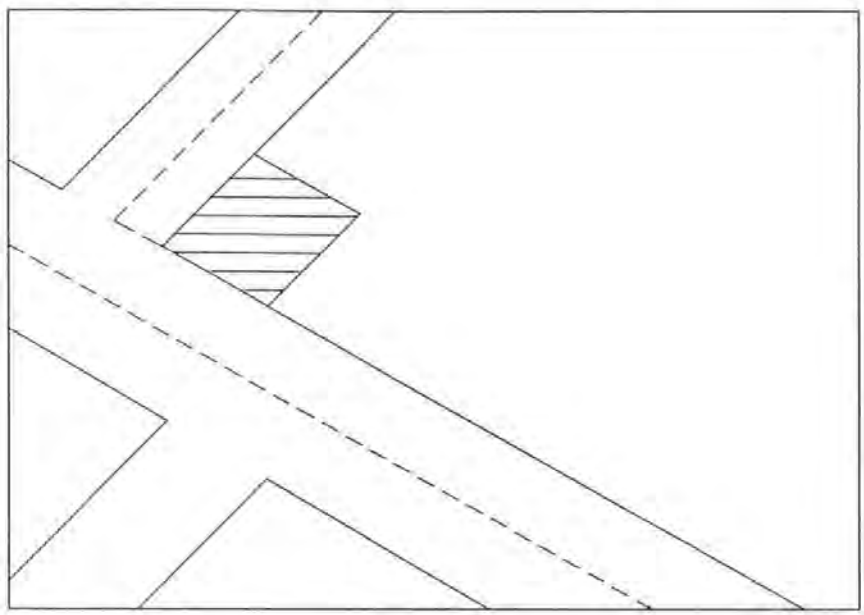
Figure 72 Penetration Domain

The Building domain takes all buildings over three stories or those buildings of one and two stories that are know to operate as businesses within the area under consideration, see Figure 73. Rules that account for telephones per square metre and turnover are then used to estimate the number of telephones and the associated calling rates. Uncertainty in the input data, the associated uncertainty windows of the rules and the resultant output inferred-data are similar to the penetration domain above.

The final Customer domain takes account of named identifiable customers, see figure 74. Here it is possible to be more accurate on the telecommunication needs of each customer. However, even here uncertainty is introduced as such customers are likely to over estimate their requirements as the element of cost is rarely taken account of at the survey stage.

7.3.2 Phase 2 of the Plymouth Expert System Operation

Phase 2 of the Expert Systems operation focuses on the development of the traffic matrix, using the Heuristic Demand Algorithm. To change the telephone densities in each 250*250 metre square into a single point source of traffic and a list of destination 'sinks', two problem domains, Exchange and Traffic Matrix, operate in parallel, each taking input inferred-data from Customer, Penetration and Building domains. The Exchange domain takes a total of sixteen 250*250 metre squares and groups them to give a single source and sink node of one square kilometre, see figure 75. This new area is likely to be a good estimate of the smallest practicable local traffic collecting area. The rule and fact uncertainty windows within each domain now combine with those associated with the input inferred-data.



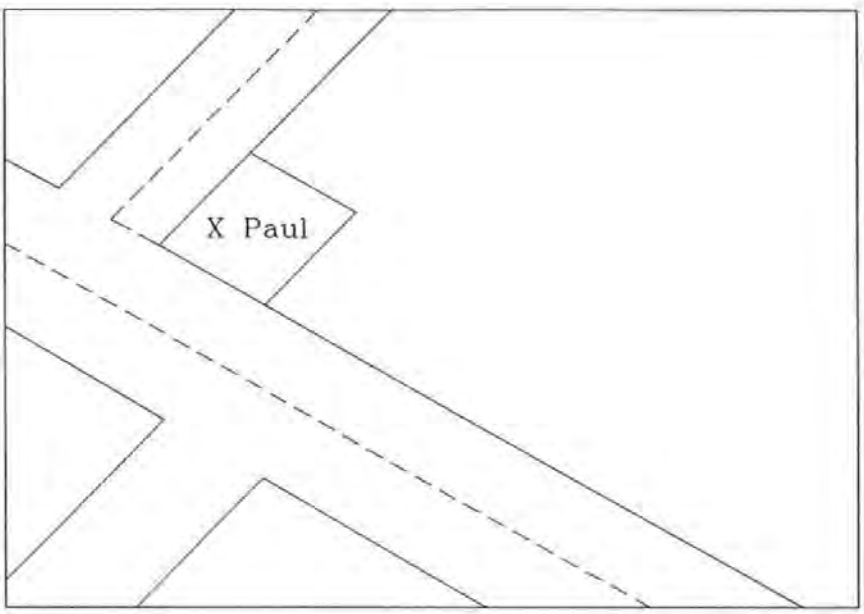
All buildings over 2 stories

- No. floors
  Area of floors
  Usage of floor - Bank
  Insurance
  Finance
  Real Estate
  Agency
  Consultancy
  Travel
  Airline
  Retail
  Construction
  Hotel
  Hospital
  Government

Each have turnover per line per sq.metre

Telephones = Area & Function & Density Function

Figure 73 Building Domain



Known users have their specific requirements entered

e.g. Paul 2 lines, calling rate 12 calls per hour,
duration 3 minutes wiith Digital Data or ISDN

Figure 74 Customers Domain

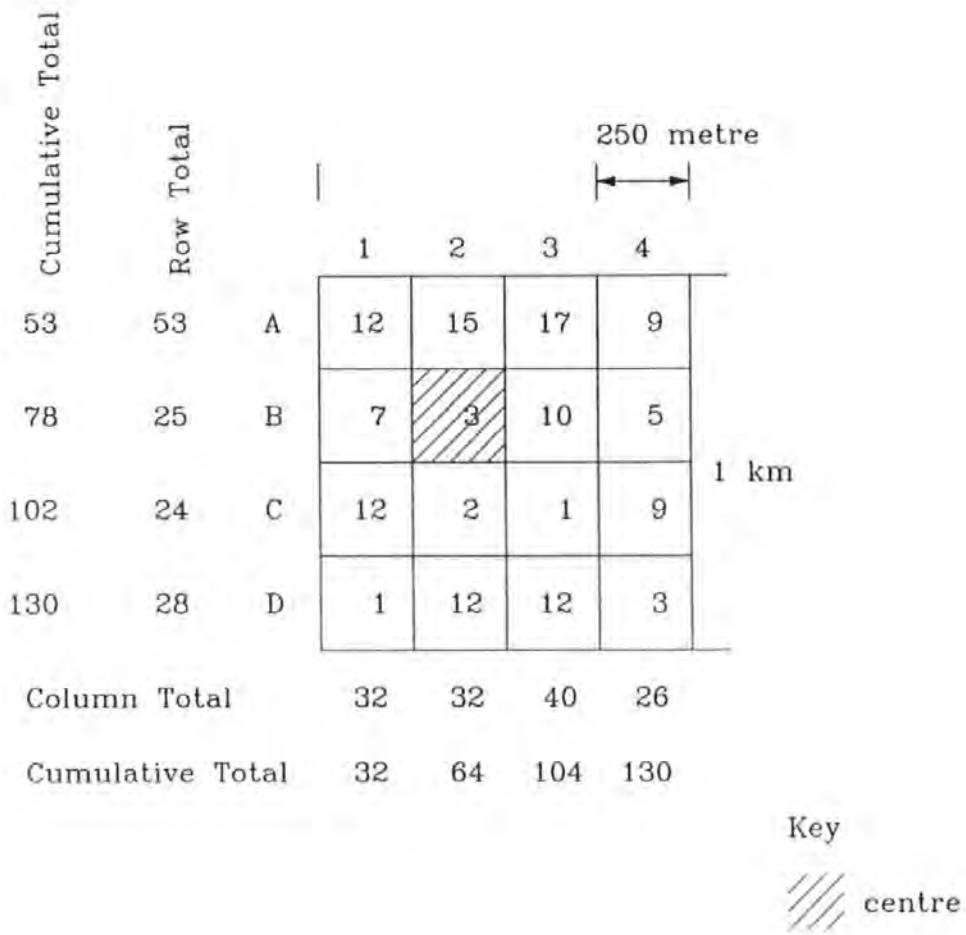


Figure 75 Smallest Traffic Collecting Area (1 km^2)

The Traffic Matrix domain attempts to generate the point source traffic matrix and communities of interest data based upon a set of heuristic rules taken from the experience of other network design problems.

Figure 76 shows one of the Plymouth Expert System displays during this phase of operation.

7.3.3 Phase 3 of the Plymouth Expert System Operation

Phase 3 of the Plymouth Expert System operation concentrates on identifying the node locations and developing the node traffic collection areas using the Node Location Algorithm within the Exchange sub-problem domain. A series of problem domains now come in to play: *Graph* develops the interconnecting link network. It initially calculates that minimum spanning tree to interconnect the source and sinks, boundary conditions, using the Link Networking Algorithm, ensuring that the maximum number of hops is kept to four. *Direct* looks at the possibilities of providing direct routes. Both domains have the aim of reducing cost and maintaining quality and grade of services. Finally, *Synchronisation* selects the location of the Primary, Secondary and Tertiary Master centres.

Figure 77 shows one of the Plymouth Expert System displays during this phase of operation.

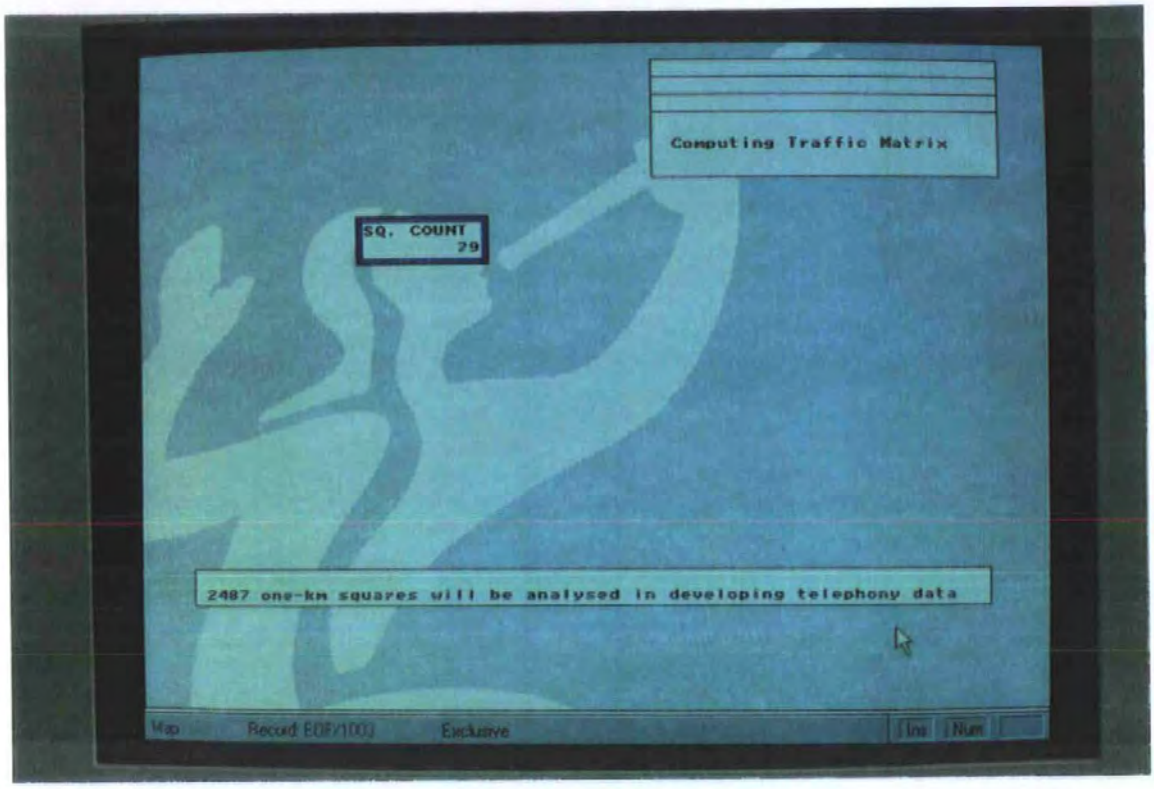


Figure 76 Plymouth Expert System Phase 2 Display (1)

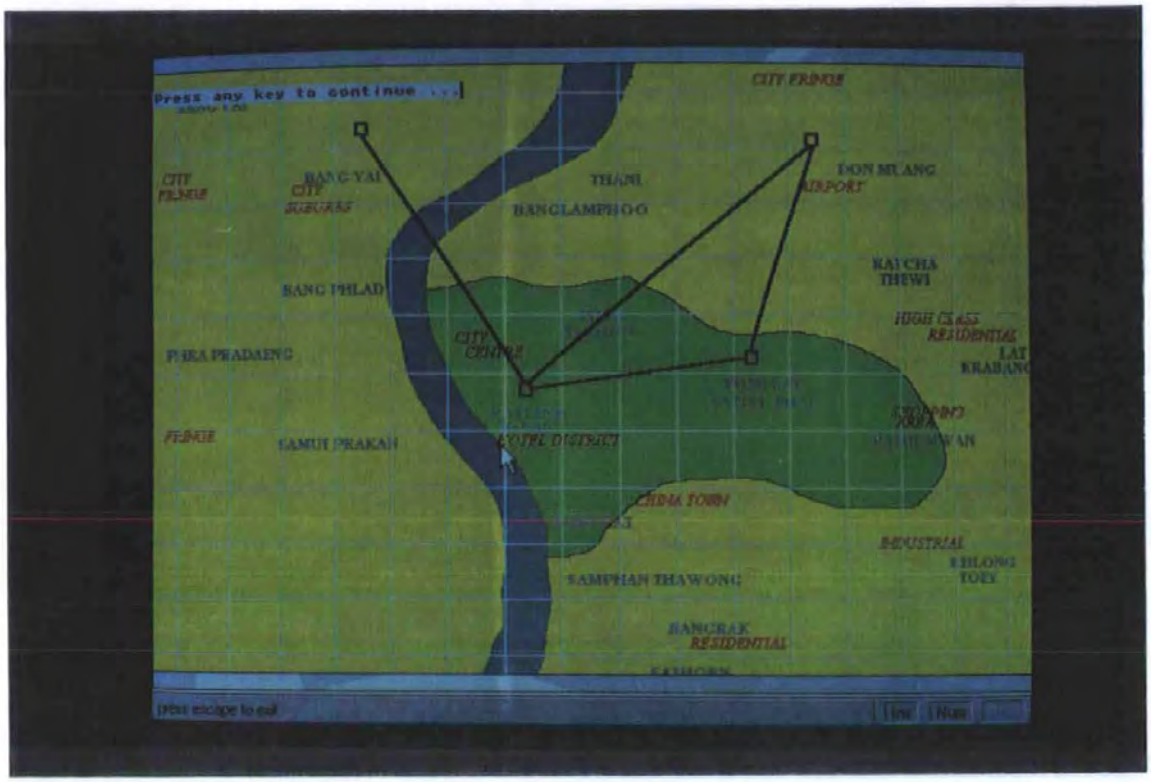


Figure 77 Plymouth Expert System Phase 3 Display

7.4 Plymouth Expert System Validation

To validate the output of the Plymouth Expert System, a comparison is made with a number of network designed in the past by BT Consultants. One example is the Thailand Case Study, outlined in Appendix 2. The Thai study not only summarises the results of the BT Consultants investigations but also that of the CCITT and JICA; and for completeness, international comparisons are also included. This example is included because it was a network designed by BT Consultants from a 'green-field' situation using traditional design techniques. The results were accepted by the BT Board and as such can form a bench-mark for the validation of the Plymouth Expert System output.

7.4.1 The Heuristic Demand Algorithm

Whilst the Plymouth Expert System is able to predict a number of data types at a micro level, e.g. traffic densities, traffic collecting areas, etc., the total number of customers was used to check the validity of the Heuristic Demand Algorithm. The results of the seven different demand assessments are reproduced below.

The figure shows that the Heuristic Demand Algorithm produces results that are compatible with those of both the Japanese and BT assessments. The main difference is the comparison the three make with the Thai Telco forecast; it is not known on what basis the Thai study was made, however, it was likely to be over optimistic to attract additional funds from their government, see Appendix 2.

1. International Comparisons,
2. BT Econometric,
3. CCITT Socio-Economic,
4. BT 'Bottom Up',
5. JICA Study,
6. Telephone of Thailand Company Forecast,
7. Plymouth Expert System.

year	1	2	3	4	5	6	7
1990	0.187	0.9	0.9	1.176	-	1.44	-
1991	-	-	-	-	-	-	-
1992	-	-	-	-	1.467	1.70	1.4
1993	-	1.3	1.1	-	-	1.86	-
1994	-	-	-	-	-	-	-
1995	0.395	1.66	1.3	1.49	-	-	-
1996	-	-	-	-	-	-	-
1997	-	2.0	1.5	-	2.12	2.53	-
1998	-	-	-	-	-	-	-
1999	-	-	-	-	-	-	-
2000	0.720	2.45	2.2	2.28	2.6	3.40	-

Forecast for Bangkok

7.4.2 The Node Location Algorithm

A comparison of the node locations chosen by the Plymouth Expert System and that of BT and the Thai Telco is shown in figure 78. Clearly, there is no comparison with the Thai Telco requirements and that produced by the Expert System and BT Consultants. This was due to the different objectives of the two teams. BT was looking to provide a commercially viable network that was cost effectively matched to demand. The Thai Telco was looking to provide ubiquitous service irrespective of demand to augment an existing network. It is difficult to say which approach is better, as they are based upon different philosophies. Comparing the BT and Plymouth Expert System results the number of nodes and their locations are, on the whole, similar; the BT results having the addition of remote peripheral units radiating from the nodes. BT chose locations on a set of subjective premises, i.e.

availability of accommodation, proximity of important people, etc., that were not available to the Plymouth Expert System: in spite of this fact, the overall cost of the three solutions is similar. The results produced by the Plymouth Expert System have proven to be sufficient that BT Research and Development Department are considering adopting the software, with modifications. Thus it can be said that the Node Location Algorithm has been successful in finding the optimum locations of nodes and their traffic collection areas.

THAI TELCO	BT CONSULTANCY	PLYMOUTH EXPERT SYSTEM
All existing Node Locations	Bang Yai + 3 Remotes	Bang Yai
	Kutapani + 3 Remotes	Kutapani
	Pamprat Satphani + 3 Remotes	Pamprat Satphani
	Don Maung + 3 Remotes	Don Maung

Figure 78 BT 'Vs' Plymouth Expert System Node Locations

7.4.3 The Link Networking Algorithm

The Plymouth Expert System did not produce the same topology as the BT study. This was disappointing. The main reason for this is thought to be that the calculation of the number of nodes necessary to support a given demand is rather too simplistic; by dividing the total demand by an average node capacity takes no account of the use of remote peripheral units which could be attached to the local nodes. These remotes could be quite small and numerous and hence produce a topology containing a greater

number of 'star' sub-networks. It is thought that more work in this area will produce designs better suited to public networks. However, even in its present implementation status the results produced are intuitively pleasing, i.e. they look right! The expert system, when run on private network designs does produce similar results to traditional methods.

As a result it can be said that the Link Networking Algorithm is suitable, only, for the design of private PBX networks in its current state of definition. More work is required to see if it can be modified to suit public network design.

7.4.4 Plymouth Expert System Credibility

In its current implementation status, the Plymouth Expert System can only be classed as part of a proposed prototype which shows the principle of heuristic network design with associated uncertainty windows as a guide to non-expert users of the system. What is needed is the exploitation of the ideas for commercial use. The system, in its current state, was evaluated by BT's 'Networks' and 'Expert Systems' divisions; they have seen enough evidence to warrant its further analysis and are considering some of the ideas for their incorporation into current BT' modelling tools. **[101 & 102].**

It is also likely to be of good value to other Telco's if their networks are very large and complicated, or by poor Telephone companies that have no specialised engineers or are situated in remote parts of the world.

7.5 The Uncertainty Windows

7.5.1 User Data

User data input as a 'datum & confidence' pair, as with conventional expert systems, was easy to use and caused no problems.

7.5.2 Facts and Rules

Producing the original uncertainty windows for the facts and rules proved to be more difficult than originally expected. Being objective on whether a supplier of information is reliable or not, was not easy. The author found it necessary, when first allocating an uncertainty window to a rule, to use simple profiles. If, when using these the resultant designs proved to be unsuitable when compared with designs of known solution, then the uncertainty window profiles were manually 'tuned' to give an acceptable result. With this 'feedback' process better uncertainty window profiles were obtained.

7.5.3 Inferred-Datum and Solution Planes

Many different solutions to the problem of presenting the solution plane to the user of the expert system are possible. The increasing availability of sophisticated interactive computer graphics should make it easier to experiment further with alternative approaches as a function of the user of the system, i.e. expert or novice. Indeed, as the solution plane is generated on-line on the computer screen directly

from the Plymouth Expert System output, then there is the possibility of interactively selecting and designing displays to suit the skills and special interests of the user.

Figure 79 shows three inferred-data uncertainty windows and Figure 80 shows a solution plane; both produced by the Plymouth Expert System when run on the design of the City of Bangkok study.

7.5.4 Figure of Merit

The interpretation of the uncertainty window characteristics, whilst powerful, can confront the novice with a difficult task. Indeed, a busy engineer may also only want a feel of the quality of the resultant design. To this end, each window needs an associated figure of merit based upon the computers' interpretation of the resultant solution plane. Various techniques were evaluated and these included the use of warm (red) and cold (blue) coloured masks for bad and good respectively, that fit over the window and statistical distribution analysis. The area of cuts in an overall inferred-datum uncertainty window proved to be the easiest to implement and interpret. However, taken alone, the total area of the window cannot be satisfactory. A left-skewed cut distribution can have the same area as a right cut, which means that both highly certain and highly uncertain values have the same area. With the use of a mask, the author was able to bias the distribution so that attenuation of the lower certainties that comprise the window was at its greatest. No analysis was conducted on the form the mask should take, in the current implementation is in the form of a triangle.

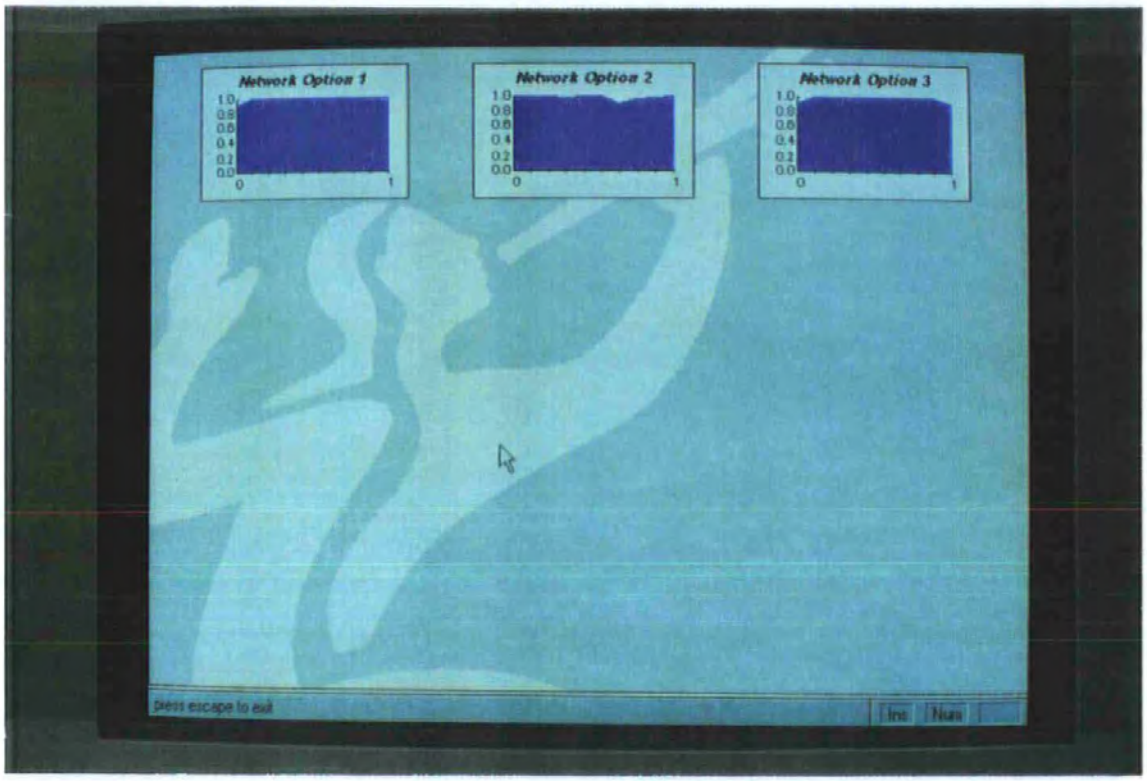


Figure 79 Three Inferred Data Windows

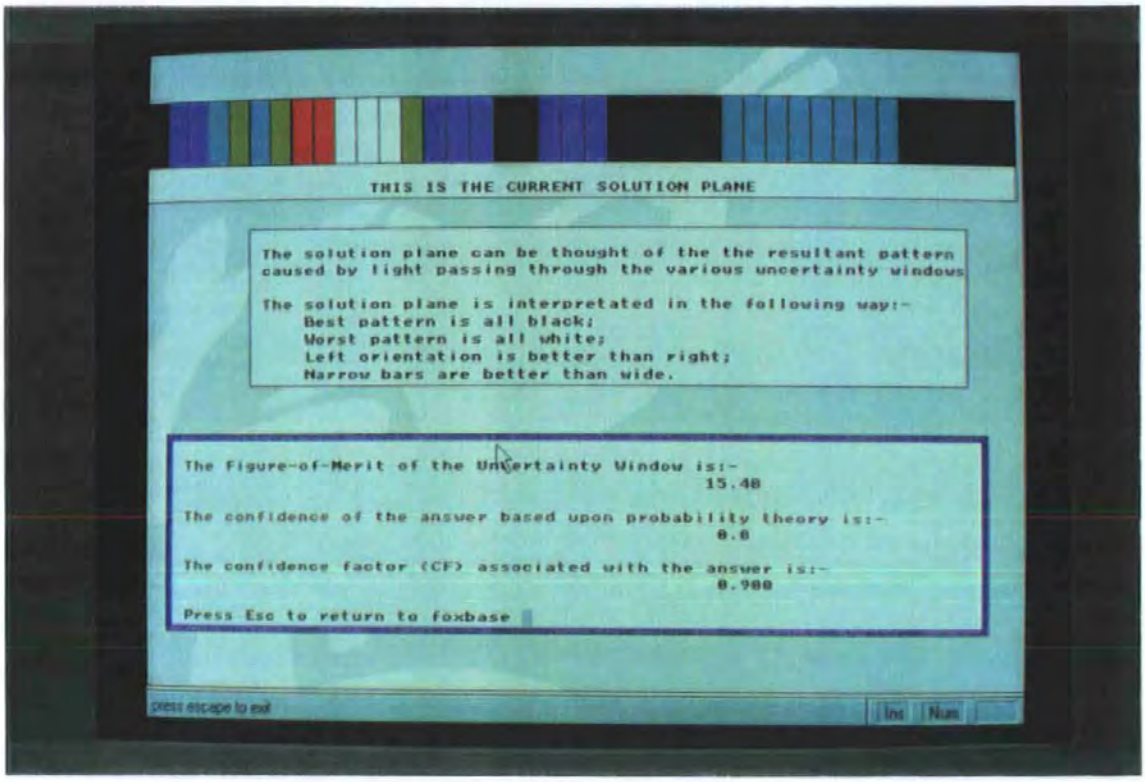


Figure 80 Plymouth Expert System Solution Plane Display

The results show that the greater the area, and hence the Figure of Merit number, the better the solution. However, extensive use of the Figure of Merit is not recommended as the intrinsic richness of the window is lost and the user is in effect reverting to current methods of dealing with uncertainty.

7.5.5 Feedback

The final design comes with its associated combined uncertainty window. It should be possible to back-track, first to the windows that made up this profile, and, ultimately to the data rules and facts that contributed to the solution plane. Thus if the output design is shown to be too unreliable in terms of the resultant solution plane profile, it is possible to focus on those parts where uncertainty is greatest and allow the designer to return to those inputs that contributed the greatest uncertainty to the design. Also if the overall attenuation is a result of the use of a significant number of unknowns the path may be re-traced. It is necessary to try to increase the profile of the solution plane such that less reliance is placed upon internal stereo-typical data.

7.5.6 Weaknesses in the Uncertainty Windows Technique

The uncertainty window process, as implemented, does have a number of weakness that needs to be overcome, two of the most significant are 'conflict resolution' and 'interpretation'.

It is very important to have control over which rule is used at any time and of the sequence in which they are operated. These factors will be dependent to a large

extent on the conflict resolution method used by the control structure but they are also influenced by the way the knowledge base has been constructed. Whilst the conflict resolution method used should not be considered above more important features such as the knowledge representation formalism or the mode of inference, it is important for anyone viewing the knowledge base to understand why rules are used when they do.

Conflict resolution is found in two distinct operational areas of the expert system: rule conflict and solution conflict. Effectively the first is on the micro scale, whilst the second is on the macro scale.

In the former, during one cycle of the inference process, it is most likely a common occurrence for more than one rule to be selected for use: one rule is chosen among this 'conflict set' by a process of 'conflict resolution'. The Plymouth Expert System overcomes this problem by choosing the rule whose uncertainty window has the highest priority; however, priority can be defined in different ways! The author has chosen, because it is both intuitively pleasing and simple to implement, a priority based upon relevance, i.e. depth of cut on the y axis of the window. A more logical method might have been to prioritise use of rules, conditional on the consequence of their effect on other rules, and eventual solution. To implement such a conflict resolution process should be possible in Prolog as the language is able to track which rules have been used in the inference process.

The second area where conflict resolution is met is during the analysis of one problem domain where a likely occurrence for more than one solution will be found by the expert system. Here the resolution process is less obvious. A simple solution would

be to present the user with the uncertainty windows associated with the various solutions and ask for guidance on which to choose. To automate the process the computer must decide for itself which window has the better profile: the current implementation uses the 'figure of merit' process, i.e. areas of cut after they have been modified by a format mask. There must be a better way of selection based upon the graphical image direct but one has yet to be found. However, it is easy to imagine that with parallel processor computer system, the windows could be cut into a number of slices equal to the number of processors and the comparisons made between the similar slices in the conflicting windows.

The second area of weakness is in the interpretation of the uncertainty windows and their associated solution planes by non-expert users of the expert system.

7.5.7 Summary of the Uncertainty Windows

Current uncertainty calculus is not suited to the solution of complex problems such as network design. They are singular in dimension and do not facilitate feedback to improve the design. Facts, rules and data need to be treated differently. The uncertainty window is able to handle these different characteristics. The windows also present the user with a view of the whole uncertainty picture that allows for more effective perceptional enhancement of the solution.

An uncertainty window is, by definition, a multi-dimensional representation that explicitly shows information about topological and geometric relations among the components of the problem: current methods of dealing with uncertainty representation do not.

The use of uncertainty windows is useful in speeding the process of :

searching data;

recognition of relevant information;

drawing inferences from that information.

An uncertainty window also helps in finding the solution to a 'search' problem by guiding the user to areas of high uncertainty or areas for further investigation. The eye can guide the computers' attention to the current or desirable location.

'Recognition' can be strongly affected by what information is explicit in a representation and what is only implicit. In the current methods of dealing with uncertainty the perceptional work is explicit and extensive; in uncertainty windows it is automatic, implied, and easy.

'Drawing effective inferences' is dependant upon the expert system users ability to enhance, perceptually, the data provided as a solution to a problem. In an uncertainty window, this enhancement is done by a process that is computationally very cheap, i.e. drawing and viewing. Thus the interpretation is passed from the computer to the user who is given enough information from which to draw conclusions, the computer screen effectively presenting the user with the whole uncertainty picture of a solution.

7.6 Conclusions

The goal of the research was to produce a prototype expert system suitable for telecommunication network design. An expert system approach is one which allows

designs to be made by non experts in the field who have incomplete data or poor statistics. The Plymouth Expert System covers the generation of traffic data, determines the optimum node locations and the optimum interconnecting trunk network. The approach taken was to evaluate and distil the process and rules used by experts in the field of telecommunication design based at British Telecommunications Consultancy department.

The achievements of the research are as follows.

First the analysis of existing methods of telecommunication network design was performed and their limitations highlighted. In particular four areas were shown to be worthy of further investigation; data collection and its processing into a traffic matrix, the location of nodes and their traffic collection areas, the optimum interconnecting trunk network and dealing with uncertainty.

Secondly, a new method of developing a traffic matrix, suited to an expert system approach, was found called the Heuristic Demand Algorithm. Both macro and micro economic methods are employed and rules of thumb used in situations of missing or vague data. The resultant traffic matrix has no community of interest information associated with it and it is thus classified as a point source traffic matrix.

Thirdly, using the point source traffic matrix and a technique called matrix erosion, a new algorithm, called the Node Location Algorithm, was developed and facilitated the location of the nodes and their respective traffic collection areas. The results of a design using the algorithm for a minimum cost distribution of nodes based upon access costs is shown.

Fourthly, having found the node locations the interconnecting trunk network is designed using a new algorithm called the Link Networking Algorithm. This is based upon graph theory and constructing a minimum spanning tree. The constraint on the tree's construction is that there are a maximum of four hops in an end to end connection.

Fifthly, a new method of dealing with uncertainty called Uncertainty Windows was developed. The Windows work because of their ability to maintain the detail of their construction and as such those areas of greatest weakness in a resultant design can be highlighted for further investigation.

The practical implementation of the Plymouth Expert System is currently only part of a prototype and lacks the documentation necessary for its commercial exploitation, at present.

7.6.1 Further Work

Whilst the Plymouth Expert System produces good results, it still needs development if it is to have commercial usage, including:

the change of the design based upon questions? For example, it would ask the question is the site available, can we get the site, can we get adjacent site?

feedback mechanisms to be fully automated;

changes to the conflict resolution mechanism;

the enhancement of the Node Location Algorithm to deal with remote peripheral units.

A very advanced version would allow the on-line connection of call loggers giving direct access to the raw data collected on existing networks and providing optimal designed output for the current traffic demands and profiles.

It is recommended that the following work be continued in the near future, taking both the Plymouth Expert System and the Uncertainty Windowing techniques further to completion.

First, the expert system should be broken down into its modules, each forming the basis of a single expert system. For example, the Heuristic Demand Algorithm can form the input to traditional computer design packages where the traffic matrix is not available, either because of inadequacies in available statistical data or uncertainty of customer locations. The algorithm will need to be developed so that community of interest information is available as required by such packages.

Secondly, the Node Location Algorithm can form the basis of a telecommunication-network design-package where transmission cost becomes insignificant within the cost of the whole network, i.e. in a high bit rate optical fibre SDH network. It will be necessary to improve the algorithm to take account of the differing types of nodes found in practice, including remote peripherals and concentrators.

Thirdly, the Uncertainty Window solution-plane can be enhanced by giving the user of the expert system the ability to use the computer mouse, to 'click' on to a portion of

the solution window whose uncertainty is deemed unacceptable. This would result in a list of the Rules, Facts and Data being displayed which were used in the compilation of that portion of the window and data, etc. and the information to be enhanced to achieve a better uncertainty profile.

Finally a further area for investigation is the development of a mathematical model of the Uncertainty Windows technique. This will not only provide for a sounder base for their adoption in other expert systems, but also allow for their interpretation by the computer to possibly become more precise.

REFERENCES

1. Rudge, A.W.: "I'll Be Seeing You", IEE Review, Vol.39, No.6, p.235-238, **1993**

2. Hausen-Tropper, E.: "An Application of Learning Algorithms to Telecommunications Networks", 6th Int. Workshop on Expert Systems, Vol.2, p.869-881, **1986**

3. Bylander, T.: "Tractability and Artificial Intelligence", J. Expt. Theor. Artif. Intell., Vol.3, p.171-178, **1991**

4. Ray, G.P.: "Computer Network Analysis and Optimisation", PhD Thesis, University of Plymouth, **1993**

5. Grout, V.M.: "Optimisation Techniques for Telecommunication Networks", PhD Thesis, Plymouth Polytechnic, **1988**

6. Bahl, L.R. and Tang, D.T.: "Optimisation of Concentrator Locations in Teleprocessor Networks", Symp. Comp. Comms. Net. & Teletraffic, Polytechnic Institute of Brooklyn, p.355-362, April 4-6, **1972**

7. Little, J.D.C., Murty, K.G., Sweeney, D.W. and Karel, C.: "An Algorithm for the Travelling Salesman Problem", Op. Res., Vol.11, p.972-989, **1963**

8. Chandy, K.M., and Russell, R.A.: "The Design of Multipoint Linkages in a Teleprocessing Tree Network", IEEE Trans. Comp., Vol.C21, No.10, p.1062-1066, **1972**

9. Kruskal, J.B.: "On the Shortest Spanning Subtree of a Graph", Proc. American Math. Soc., Vol.7, p.48-50, **1956**

10. Prim, R.C.: "Shortest Connection Networks and some Generalisations", Bell Sys. Tech. Jour., Vol.36, p.1389-1401, **1957**

11. Esau, L.R. and Williams, K.C.: "A Method for Approximating the Optimal Network", IBM Sys. Jour., Vol.5, No.3, p.142-147, **1966**

12. Grout, V.M., Sanders, P.W. and Stockel, C.T.: "Practical Approach to the Optimisation of Large Switching Networks", Comp. Comms., Vol.10, No.1, p.30-38, **1987**

13. Bear, D.: "Principles of Telecommunication-Traffic Engineering", (Peter Peregrinus Ltd.), Chapter 3, **1976**

14. Lee, C.Y.: "Analysis of Switching Networks", Bell Sys. Tech. Jour., Vol.34, Nov., p.1287-1315, **1955**

15. Littlechild, S.C.: "Elements of Telecommunications Economics", (Peter Peregrinus Ltd.), Chapter 11, **1979**

16. Caudill, M.: "The Possibilities of Probabilities", A.I. Expert, Vol.8, No.3, p.28-31, **1993**

17. Kai-li, K.: "Expert Systems in Telecommunications Network Planning and Design", Proc. 1st Int. Con. App. A.I. in Eng., Vol.2, p.1161-1164, **1986**

18. Roth, P.F., Mohammad, I. and Mouftha, H.T.: "Simulation: a Powerful Tool for Prototyping Telecommunications Networks", Simulation, Vol.58, p.78-82, **1992**

19. Mantelman, L.: "AI Carves Inroads: Network Design, Testing and Management", Data Comms., p.106-123, July, **1986**

20. Ferguson, I.A. and Zlatin, D.R.: "Knowledge Structures for Communications Networks Design and Sales", IEEE Nets.(USA), Vol.2, No.5, p.52-58, **1988**

21. Tange, I.: "General Considerations by CCITT Working Party XIII/2 on the use of Computers for Network Planning", Tele. Jour. ITU, Vol.38, No.11, p.737, **1971**

22. Gupta, V.P.: "What is Network Planning?", Vol.23, No.10, p.10-16, **1985**

23. Penrose, R.: "The Emperor's New Mind", (Oxford University Press), p.138-146, **1989**

24. Kehoe, L.: "White Collar Roberts go to Work", Financial Times, p.6, August 5, **1986**

25. Gross, D.: "Applications of AI Technology in Communications", Expert Systems(UK), Vol.5, No.3, p.248-251, **1988**

26. Kamimura, K. and Nakajima, I.: "Expert Systems Supporting Private Network Design", NTT R&D (Japan), Vol.38, No.11, p.1377-1386, **1989**

27. Abate, J.E., Butterline, E.W., Carley, R.A., Greendyk, P., Montenegro, A.M., Near, C.D., Richman, S.H. and Zampetti, G.P.: "AT&T's New Approach to the Synchronisation of Telecommunication Networks", IEEE Com. Mag., Vol.27, No.4, p.35-45, **1989**

28. IEE Colloquium on "Reasoning Under Uncertainty", London, IEE Digest No.086, **1990**

29. Lees, C.: "Defining Expert Systems", Telecommunications, Vol.23, No.3, p.65-71, **1989**

30. Salasoo, A.: "Initiating Usable Methods with a New Engineering Design Tool", SIGCHI Bulletin, Vol.23, No.1, p.68-70, **1991**

31. Solo, A. and Hamalainen, P.: "Seteli: The Strategy Expert for Telecommunications Investment", IEEE Expert, Vol.15, No.5, p.14-22, **1990**

32. Descartes, R.: "Meditations on the First Philosophy (I and II)", **1637** (Reprinted Penguin)

33. Russell, B.: "The Problems of Philosophy", (OUP), p.1-4, **1959**

34. Russell, B.: "An Inquiry into the Meaning and Truth", (Allen and Unwin), **1940**

35. Ayer, A.J.: "The Problem of Knowledge", (Penguin Books), Chapter 2, **1956**

36. Whiteley, C.H.: "Epistemological Strategies", Mind, Vol.78, p.25-33, **1969**

37. Popper, K.R.: "The Logic of Scientific Discovery", (Basic Books), **1959**

38. Bayes, T.: "An Essay Towards Solving a Problem in the Doctrine of Chances", Phil. Trans. Royal Soc., London, Vol.LIII, p.370-418, **1763**

39. Aleliunas, R.: "An Inquiry into Computer Understanding", Comput. Intell., Vol.4, p.67-69, **1988**

40. Aristotle: "The Politics", **384 B.C.**
[Translated, Sinclair, T.A., (Penguin Classics), p.64-65, (1253b23), 1962]

41. Lovelace, Countess of: "Translator's Notes to M. Menabrea's Memoir on Babbage's Analytical Engine", Scientific Memories (ed. Taylor, R.), Vol.3, p.691-731, **1842**

42. Wiener, N.: "Cybernetics - Or Control and Communication in the Animal and the Machine", (MIT Press), **1948**

43. von Neumann, J.: "Report on the EDVAC", University of Pennsylvania, **1945**

44. Jefferson, G.: "The Mind of Mechanical Man", Brit. Med. Jour., Vol.10, p.1105-1110, **1949**

45. Turing, A.M.: "Computing Machinery and Intelligence", Mind, Vol.LIX, No.236, p.433-460, 1950

46. Epstein, R.: "The Quest for a Thinking Computer", AI Mag. (USA), Vol.13, No.2, p.80-95, 1992

47. Minsky, M.: "Steps Towards Artificial Intelligence", Proc. IRE, Vol.49, No.1, p.8-30, 1961

48. Feigenbaum, E.A.: "Expert Systems in the 80s", Infotech Report-Machine Intelligence, p.219-229, 1981

49. Clark, D.A.: "Numeric and Symbolic Approaches to Uncertainty Management in AI", Artif. Intell. Rev.(UK), Vol.4, No.2, p.109-146, 1990

50. Cohen, P.: "Heuristic Reasoning about Uncertainty", (Pitman), p.184, 1985

51. O'Leary, D.E. and Kandelin, N.A.: "Validating the Weights in a Rule-based Expert System", Int. J. Expert Syst. App., Vol.11, No.3, p.253-279, 1988

52. Cartes-Rello, E and Golshai, F.: "Uncertainty Reasoning using the Dempster-Shafer Method: An Application in Forecasting and Marketing Management", Expert System(UK), Vol.7, No.1, p.9-18, 1990

53. Hughes, C. "The Representation of Uncertainty in Medical Expert Systems", Med. Info., Vol.14, No.4, p.269-279, 1989

54. Shiraishi, N., Furuta, H., Umano, M. and Kawakami, K.: "Knowledge Based Systems for Damage Assessment", Proc AI Civil and Structural Engineers, London, p.211-216, **1989**

55. Zwick, R. and Wallsten, T.S.: "Combining Stochastic Uncertainty and Linguistics Inexactness", Int. J. Man-mach. Studies, Vol.30, No.1, p.69-111, **1989**

56. Gaag, L.C.: "Different Notions of Uncertainty", Int. J. Man-mach. Studies, Vol.33, No.5, p.595-606, **1990**

57. Murphy, S.V. and Kandel, A.: "Fuzzy Sets and Typicality Theory", Info. Sci., Vol.51, No.1, p.61-93, **1990**

58. Guan, J., Bell, D.A., Pavlin, J. and Lesser, V.R.: "Dempster-Shafer Theory", IEE Colloquium on Reasoning Under Uncertainty, Digest No.086, p.6/(1-3), **1990**

59. Sheridon, F.K.J.: "A Survey of Techniques for Inference Under Uncertainty", Art. Intell. Review, Vol.15, No.1 & 2, p.89-119, **1991**

60. Spiegelhalter, D.J.: "Probabilistic Reasoning in Expert Systems", J Math. Man. Sci., Vol.9, No.3 & 4, p.191-210, **1989**

61. Henkind, S.J. and Harrision, M.C.: "An Analysis of Four Uncertainty Calculi", IEEE Trans. Sys. Man. & Cybern., Vol.18, No.5, p.700-714, **1988**

62. Fox, J.: "Expertise in Man and Machine", IEE Colloquium on Application of Knowledge Based Systems, p.2/1-4, 1982

63. Averkin, A.N. and Nguyen, M.K.H.: "Utilising Fuzzy Relations in Knowledge Representation", Sov. J. Computing Sci. (USA), Vol.28, No.2, p.55-68, 1990

64. Szolovits, P. and Pauker, S.G.: "Categorical and Probabilistic Reasoning in Medical Diagnosis", Art. Intel., Vol.11, p.115-144, 1978

65. Miller, R.A., Pople, H.E. and Myers, J.D.: "An Experimental Computer Based Diagnostic Consultant for General Internal Medicine", N. Engl. J. Med., Vol.307, No.8, p.468-476, 1982

66. Shortliffe, E.H. and Buchaman, B.G. "A Model of Inexact Reasoning in Medicine", Math. Bio. Sci., Vol.23, p.351-379, 1975

67. Brooks, R. and Helser, J.: "Some Experience with Transferring the MYCIN System to a New Domain", IEEE Trans Pattern Analysis and Machine Intelligence, PAMI-2, p.477-478, 1980

68. Hajek, P. and Havranek, T.: "Combining Functions for Certainty Degrees in Consulting Systems", Art. Intell. & Info. Sys. of Robots, (Proc. 4th. I.T. Conf.), Smlolenice, Czechoslovakia, p.41-47, 1987

69. Negoita, C.V.: "Expert Systems and Fuzzy Systems", (Benjamin/Cummings USA), Chapter 5, 1985

70. Zadeh, L.A.: "A Theory of Approximate Reasoning", Mach. Intell., Vol.9, p.149-194, 1979

71. Adlassnig, K.P., Kolarz, G. and Scheithauer, W.: "Present State of the Medical Expert System CADIAG-2", Methods Inf. Med., Vol.24, p.13-20, 1985

72. Adlassnig, K.P.: "A Survey of Medical Diagnosis and Fuzzy Subsets", Approximate Reasoning in Decision Analysis, (Gupta and Sanchez), North Holland: Amsterdam, p.203-217, 1982

73. Fieschi, M., Joubert, M., Fieschi, D., Botti, G. and Roux, M.: "A Program for Expert Diagnosis and Therapeutic Decision", Medical Informatics, Vol.8, No.2, p.127-135, 1983

74. Gaines, B.R.: "Foundations of Fuzzy Reasoning", Int. J. Man-Mach., Vol.8, p.623-668, 1976

75. Kahneman, D., Slovic, P. and Tversky, A.: "Judgement Under Uncertainty: Heuristics and Biases", (Cambridge University Press), p.12, 15, 141, 231-237 & 357, 1982

76. Hink, R.F. and Woods, D.L.: "How Humans Process Uncertain Knowledge", AI Mag(USA), Vol.8, No.3, p.41-53, 1987

77. Magill, W.G.W. and Leech, S.A.: "Uncertainty Techniques in Expert Systems Software", Decis. Support Sys., Vol.7, No. 1, p.55-65, 1991

78. Castillo, E. and Alvarez, E.: "Uncertainty Methods in Expert Systems", Micro Civil Eng.(USA), Vol.5, No.1, p.43-58, 1990

79. Garbolino, P.: "Bayesian Theory & AI", Int. J. Man-mach. Studies, Vol.27, No.5 & 6, p.729-742, 1987

80. Ibrekk, H. and Morgan, M.G.: "Graphical Communication of Uncertain Quantities to Non-Technical People", Risk Ana., Vol.7, No.4, p.519-529, 1987

81. Larkin, J.H. and Simon, H.A.: "Why a Diagram is (Sometimes) Worth Ten Thousand Words", Cog. Sci., Vol.11, No.1, p.65-99, 1987

82. Cleveland, W.S. and McGill, R.: "Graphical Perception and Graphical Methods for Analysing Scientific Data", Science, Vol. 229, p. 828-833, 1985

83. McDermott, J.: "R1: A Rule-Based Configure of Computer Systems", Art. Intell., Vol.19, p.39-88, 1982

84. Quinlan, J.R.: "Fundamentals of the Knowledge Engineering Problem", Infotech Report - Machine Intelligence, p.293-305, 1981

85. Charniak, E.: "AI Programming", (Lawrence Erlbaum), 1980

86. Kowalski, R.A.: "Logic for Problem Solving", (North Holland Inc.), 1979

87. Clocksin, W.F. and Mellish, C.S.: "Programming in PROLOG", (Springer-Verlag), **1981**

88. "GAS 3" and "GAS 5" Handbooks, (CCITT), **1990**

89. "International Financial Statistics", (International Monetary Fund), Vol.XLVI, **1993**

90. "Yearbook Of Public Telecommunication Statistics", Edition 20, (CCITT), **1993**

91. HM Treasury: "Offer for Sale - British Telecommunications PLC", (S.G.Warburg & Co. Ltd.), **1993**

92. Reynolds, P.L.: "System Consultancy Report", (Business Press International Ltd.), **1983**

93. Reynolds, P.L.: "Group (UK) Telecommunications Network Design Study: Main Report", Report No. Tel 00184, (British Petroleum), **1984**

94. Reynolds, P.L.: "PRISM: A Network Consultancy", (British Telecommunications PLC), **1984**

95. Reynolds, P.L.: "Potential use of MDC's Telecommunications Network to Support Integrated Services Resale", (McDonald Douglas Corporation), **1985**

96. Reynolds, P.L.: "Navy Department ISDN - Network Design Study" (Restricted), (Ministry of Defence), **1986**

97. Reynolds, P.L.: "A Business Overlay Network for Maharastra State - India", (British Telecommunications PLC), **1989**

98. Reynolds, P.L.: "A Traffic Matrix for Thailand", (British Telecommunications PLC), **1990**

99. Berry, L.T.M.: "An Explicit Formula for Dimensioning Links Offered Overflow Traffic", Aust. Telecom. Review, Vol.8, Part 1, p.13-17, **1974**

100. Pratt, C.W.: "The Concept of Marginal Overflow in Alternative Routeing", Aust. Tele. Res., Vol.1, p.76-82, **1967**

101. Smith, R.: Private Referee Report, July, **1993**

102. Dufour, I.: Private Correspondence, BT Reference - IGD838/AC, August, **1993**

APPENDIX 1

CASE STUDY ONE

The Heuristic Synchronisation Problem

Need for Network Synchronisation

Pulse Code Modulation (PCM) transmission, and all derived functions such as digital switching, must be synchronised. PCM transmission systems are kept in synchronism by means of clock and frame alignment. Here the de-multiplexer extracts information (frame alignment bits code pattern) from the incoming bit stream to derive an identical clock as the transmit end clock.

In a telecommunication network, the PCM transmission systems radiate from a digital switch. The switching rates within the digital switch may, typically be governed by the frequency of a local quartz crystal oscillator of the timing unit (TU) of the digital switch. This is a reliable frequency source, but there will be minor differences between the natural frequencies of the TUs in a network because the natural frequency of their oscillators may drift from nominal (e.g. with age).

Thus, the frequency of a network of synchronised oscillators will drift unless the network is locked (synchronised) to a frequency source that has an acceptable margin of accuracy and reliability. The term synchronised, as used here, refers to an arrangement for operating digital switching systems at a common (synchronised) clock rate.

If the digital switching clocks are not synchronised, the information rate of a signal received at a switch could not process the information and re-transmit it. This would

result in information being lost (if the input rate was faster than the receiving switch local clock) or repeated (if it was slower).

The deletion/repetition of a single PCM frame is called a 'slip'. A slip represents distortion of information, normally eight bits of each channel. The resulting service impairment is far more serious for data customers (causing errors in data signal) than for voice customers (causing clicks and distortion). This arises since the voice is highly redundant whereas data is not.

Network Hierarchical Synchronisation

CCITT's GAS 3 handbook introduces Network Synchronisation Planning (NSP) outlines the various techniques for synchronisation.

A hierarchical synchronisation plan is adopted for most public digital networks: private digital networks normally adopt the same basic philosophy.

The BT digital switches are divided into 4 hierarchical levels, level 1 being the highest in the synchronisation hierarchy, and level 4 the lowest.

To ensure that the mean frequency of the digital switching clocks is the same, these are by way of the 2.048 Mbit/sec digital line links. Synchronisation links between the switches will be either unilateral (with control being effective at one end only), or bilateral (with control being effective at both ends). It will be seen from the figure that there are unilateral links between nodes (switching centres) at different levels (with the effective end at the lower level node) and bilateral links between nodes at the same level.

A single level 1 node (one of the main trunk switches or one of the international switches) with an operationally secured (triplicated) caesium (atomic) clock of accuracy 1 and 10^{11} will function as the national reference switch. This caesium frequency standard is known as the 'National Synchronisation Reference Clock'.

In the hierarchical arrangements, a node can only receive the synchronisation reference signal from another node, distinct from itself, which contains a clock of superior or at least equivalent quality. The flow of synchronisation reference is mainly down the hierarchy, with some horizontal excursions, but never upwards.

Synchronisation utility (SU) equipment at each digital switch is connected to the synchronisation line links, i.e. nominated 2.048 Mbit/sec PCM systems. The SU derives information from the synchronisation link to control the frequency of the switch local clock by 'phase-locked-loop', and thus maintains it in synchronism with the rest of the network.

While the clock accuracy at the highest level is 1 in 10^{11}, the accuracy at the lower levels will be determined by the characteristics of the phase-locked-loops. Typically, accuracy at the lower level 4 is likely to be 1 in 10^{8}.

Two items comprise the SU:

- Link Control Unit;
- Common Control Unit.

The CCITT Recommendation G703 requirement for public networks at the lowest level is 2.048 Mbit/sec ±50 parts per million (ppm). (NB: CCITT Recommendation Q513 Section 2 gives further detail of timing and synchronisation.)

Failure of network synchronisation at any node would cause free-running of the local clock at that node to cause drift and thus the plesiochronous mode of operation in the abnormal condition. This could lead to slips. (Plesiochronous - where the frequency sources are independent but with closely defined limits.)

Private Network Synchronisation

Private digital networks follow the same basic philosophy adopted for public networks in the sense that the synchronisation network is hierarchical and that a private network reference clock is required.

Since private networks are smaller than their public equivalents, the number of synchronisation levels will be more modest although a synchronisation plan could have up-to three defined masters.

It is logical to assume that the expense of operationally secured atomic clocks as the reference source as part of the private network is not justified. In this circumstance, one or two nominated main switching centres in the private network could be fed with 2.048 Mbit/sec link(s) from the public network (BT) reference clock. These nominated private main switching centre(s) then functions as the primary masters for the private network.

By heuristic methods other switches in the network are then selected as Secondary and Tertiary Masters.

Private networks use PBXs as their switching hardware and in all UK supplied switches, the SU & TU provided with such are not to the same standard as Public switches, indeed BT has developed the Equipment Synchronisation 6001B to provide for this.

The technology restricts the planner to Master-Slave techniques and allows for changes in the synchronisation plans in predetermined order, i.e. Predetermine Master Slave.

Basic Principles and Strategies

The pure Master-Slave synchronisation method is based upon each switch in the digital network being phase-locked to another in such a way that all attain the frequency of a selected master-switch called the 'Primary Master' (PM). The procedures described in this section detail a method of selecting the PM and include methods for selecting alternative synchronisation paths for change-over in case of digital link failures. (This method is sometimes named Pre-selected Alternate Master Slave.) [As a first-cut assumption all synchronisation paths are assumed to be unilateral (UL) and each switch having a hierarchical order of standby links or clocks.]

To aid the design of a Network Synchronisation Plan (NSP) the concept 'Secondary Master' (SM) and 'Tertiary Master' (TM) is used. These are switches that are normally slaved to the PM but which may change-over to an accurate local reference

clock with loss of the source synchronisation links. However, depending on the network structure, there may be other switches that will function in the same or similar manner as SM and TM and have the same clock accuracy. Hence, SM/TM are used merely as an aid to obtaining a good NSP.

Optimum NSP Designing

A NSP is considered being optimum if:

- a failure of a digital route has a minimum impact on traffic flow measured in terms of slip-rate on remaining digital links;

- it gives a more stringent requirement on 'transit switching' compared with 'local switches' measured by the probability of a 'degraded' slip rate performance level. (This is in accordance with CCITT G.822);

- it uses a minimum number of accurate local reference clocks for a given level of network performance.

A NSP is considered being resilient if:

- whenever possible, a switch has at least 2 or preferably 3 independent (with respect to availability) reference sources;

- whenever a switch has two or three digital routes as references, 2 digital links in each digital route are included in the hierarchical order of synchronisation references. (Thus, a pair of links is selected in each digital route making a

maximum of 6 digital links. It is sometimes possible to connect more links, but little is gained);

- when a switch has only one digital route as a reference, 3 digital links within that route are selected.

Existing Networks

Most existing networks are expected to change in the number of switches and most particularly, the transmission media connecting them. Also existing networks may comprise both analogue and digital routes; the digital network may have a structure that is different from the routeing network. Because of these aspects and the fact that differing switch manufacturers may not have sufficient synchronisation capabilities, the designers of the NSP normally need to apply a large amount of common sense and network evolution knowledge.

Creating the Plan - Heuristic Rule Base

This section of the report details a rule base for generation of a master-slave NSP for networks. By going through a series of steps, a plan giving a priority order of references and clock requirements is developed. It will, at a later stage, need to be enhanced to take account of satellite working.

Although the procedures described here should be applicable to networks with switches of any type, its tested validity has been restricted to BT PBX's

The NSP obtained by going through the outlined procedures will be a plan for a network topology about three to five years ahead of the present. Therefore, interim plans will also have to be developed for use during the development stage. These interim plans can be designed with the knowledge of the medium or long term NSP and the experience gained while developing it. The changes that have to be made in the life of the NSP should preferably be minor.

The procedure is organised in the following way:

- identify the digital network,

- primary Master and Star-structures,

- secondary/Tertiary Master and Link Priorities,

- allocation of final link priorities.

• Identify the Digital Network

STEP 1

Identify all existing as well as future digital switches or analogue switches with, or will have, digital transmission to another switch (i.e. output of the Plymouth Model). Make a list of all switches noting all relevant facts such as installation year, clock equipment, synchronisation capabilities, etc. On a network topology 'map', designate each switch as Transit, or Local. (For combined switches, choose the higher level).

STEP 2

Identify all existing as well as future (approximate three to five years hence) digital transmission facilities between all switches listed in STEP 1. Organise the transmission facilities into digital links and digital routes. (A digital link is a 2048 kbit/sec transmission system a digital note is a collection of digital links that criticise a common transmission object which if affected would effect all links within that route).

- **Primary Master and Star-structures**

All networks can be broken down into a combination of stars and meshes. (Meshes interconnected by stars and vice versa). Those parts of the network that exhibit a star-like structure (star-section) have very obvious synchronisation plans once the Reference node is known.

In a pure master-slave plan, the Reference node is called Primary Master.

STEP 3

- **Selection of Primary Master**

A first assumption in this step is that PM is a switch. However, it is possible to select a special clock-node as PM as long as it has a timing distribution similar to the digital routes of a transit switch.

As an aid to the selection of the Primary Master, the list below highlights some desirable properties of the required site. Of these only the last can be classified as an essential requirement. Therefore, select the node that has the most appropriate combination of the following properties:

- have a large number of digital routes,

- be a transit switch,

- have or will get connection to the Public Switched Telephone Network (PSTN),

- be a 'nucleus', i.e. the deepest node, within the network,

- have provisions for connection of accurate reference clocks.

Name this the 'PM' and draw arrows on all digital routes radiating from the PM, called 'arrowhead-routes', to all switches it is connected to.

Set a Priority P = 1 on these arrowhead routes.

STEP 4

• **Star-structures:**

All star-sections are now synchronised in the very obvious way. Emanating from the PM or a Reference node or the mesh that comprises it, allocate arrows on the digital routes in all star-sections.

On the network 'map', each digital route is drawn as one line connecting two switches. Certain distances of a digital route may be common with another digital route of another switch, e.g. a common microwave link or a cable route which further on branches off into geographically separate cables. Such common paths may be drawn on the 'map' as two (or more) parallel lines with a ring around them.

(For clarification, digital link is a link as seen from the switches synchronisation utility, e.g. a 30 channel system irrespective of any further higher order multiplexing.) Digital links that are semi-permanently connected through a switch are not a digital route of its own right they should be grouped together with the other digital links to that switch).

A list of all digital routes is made, noting all relevant facts such as installation year, physical type, length and the number of digital links to aid future planning.

Important facts should be also noted on the 'map', e.g. if a digital route has only 1 digital link. Also mark on the 'map', with a cross, those digital routes that may be inappropriate as synchronisation references.

• Secondary & Tertiary Master and Mesh Structures

This routine is structured into the following sections:

- selection of secondary master (SM),

- selection of tertiary master (TM),

- selection of the other switches within the mesh.

The digital network may be composed of (beside pure stars) several mesh-sections that have no connections between them except through star-sections or by way of the Primary Master switch. The following routine is recursive so it should analyse all switches in every mesh-section of the digital network.

A simple test to see if a switch belongs to the same mesh- section as the previously selected switch is: it belongs to a mesh if a digital route leads to a previously selected switch without crossing the master-node (normally PM) or using a star-section.

• Selection of Secondary Master

If no switch can be selected by the following routine, and go to 'Remainder' Also mesh-sections with only local switches do not have SM therefore go to 'Remainder'.

STEP 5

Select a switch, previously not selected, using the following criteria:

- Shall have an arrowhead-route from PM. Preferably 2 (diverse) arrowhead-routes from PM. If there is only one, it should preferably not run parallel with any other digital arrowhead-route from PM and the route should comprise at least two digital links.

- Shall, if possible, have a digital route to another arrow- marked switch whose arrow-route is received by a diverse way.

- Should have the smallest number of digital arrowless routes, but at least one, which may lead back to PM or to a Public switch (by way of other switches).

- Shall be able to be equipped with an accurate local reference clock.

- If ambiguity remains, consider the quality of digital links. Furthermore, a transit is preferred to a local switch.

Name this the 'SM' and draw arrows on all digital routes to all switches it is connected to (except to PM of course). If the destination switch has two diverse arrows from PM, set priority P = 2 on all outgoing arrowhead routes else set priority P = 3.

• **Selection of Tertiary Master**

If all the remaining switches in the mesh- section are local switch: ignore the following routine and go-to 'Remainder'.

STEP 6

Select a switch, previously not selected, using the following sequence of selection criteria:

- Shall have an incoming arrow from 2 switches on 2 diverse routes and at least 1 arrowless digital route. If such a switch does not exist, return to 'SM' and select one more SM within the same mesh-section as the previous SM.

- Shall have the largest number of arrowless routes, that may lead back to PM. If the number of such digital-routes is zero, then go-to 'Remainder'.

- Shall be able to be equipped with an accurate local reference clock.

- If ambiguity remains, consider route diversity, and route length. (And the quality of the routes?) Also, a transit, is preferred to a local switch.

Name this the 'TM' and draw arrows on all digital routes to all switches it is connected to (except to PM or SM of course). If it has an arrow from PM, set priority P = 2 on all outgoing arrows, else set to priority P = 3.

• **REMAINDER**

The remaining switches in the mesh-section should be synchronised using the following routine.

STEP 7

Let P=3 (start value for priorities in the routine below) &

Let N be the number of incoming arrows.

Loop while this routine can select a new switch.

Identify that or those switches, previously not selected, which fulfils the following criteria: (If no switch fulfils the criteria, it implies that you shall exit this do-while loop):

- Shall belong to the same mesh-section as the previously selected switches unless a new mesh-section shall be entered which lacks SM,

- Shall have the maximum 'N', preferably from diverse routes, and have at least 1 arrowless digital route.

The number of arrows, N, to the identified switches determines the next action:

N = 1:

If there are more than one such switch, the following may do the last selection:

- The arrow should comprise at least two digital links and they should not run parallel with any other arrowhead route,

- It should have and arrowless digital route to another arrow- marked switch whose arrow is received on a route diverse from this arrowless digital route,

- A transit switch is preferred to a local switch,

- Few arrowless digital routes is preferred to many, ;

- There is a preference for high priority arrows. (1 before 2 and so on).

Let P = P + 1

Draw arrows on the arrowless digital routes from this switch. Set priority P = P+2 on these arrows.

N>1:

If there are more than one such switch, the following may do the last selection:

- The switch with the highest priority arrow should be selected,

- Consider the diversity and quality of the routes. (It is good with 2 digital links onthe highest priority route).

Let P = old P + 1

Draw arrows on the arrowless digital routes from this switch to those switches that have less than three incoming arrows. Set priority P = P on these arrows.

End of do-while loop.

By now, this mesh-section should be completed. If switches remain in any other mesh-section with no arrows, return to SM and restart the iterative process.

• Renumbering of Priorities

The NSP will now have been developed such that each switch has one to three arrows, each with a priority number. However, it is appropriate to use 2 (or

sometimes 3) digital links in each digital route, to minimise the impact from hardware failures in equipment. Hence, a new priority order has to be established for each switch where all digital links are ordered in a straight hierarchical priority order.

It is appropriate to make a list for each switch naming the digital links and the clocks that are to be connected and their priorities to be set from command, to provide for an easy and safe installation. The following procedure applies:

STEP 8

- If the switch has only one arrow, select, if possible, 3 digital links in that digital route and set priorities 1 to 3 on these links,

- If the switch has more than one arrow, select, if possible, 2 digital links in each digital route. Starting with the highest priority digital route (smallest priority number), set priorities for all digital links in straight increasing order starting from one and up,

- Sometimes two digital routes have the same priority number and in that case it does not matter that you select first. However, it is practical to select first the one that has the shortest path to the primary master,.

- Additional, non switch-Clocks are placed last in the priority order.

As the rule base stands no account has been taken for security against faults.

- Equipment and Link Security

Any 2.048 Mbit/sec digital transmission link may be used for synchronisation control purposes: Provided that there is sufficient capacity for synchronisation information to be on each link, there is no other special requirements for synchronisation.

However, to provided security of synchronisation information a number of changes are required to the rule base.

- The failure rate of a line system is length dependent, so that links contributing to synchronisation should be as short as possible.

- They should also follow physically diverse routes when possible, for if they follow the same routes there is a possibility that a single cable fault could affect more than one link.

- Each transmission link contributing to the synchronisation system requires a LCU to compare the phase of the local clock with that of the clock at the far end of the link. Each LCU should be fail-safe to ensure that its malfunction has no adverse effects. A number of links participate in the synchronisation control at each switch, and thus failure of an individual transmission link or LCU should not cause the clock at that switch to lose synchronisation with clocks at other switches.

- The SU must be highly secure, for loss of this equipment invariably means a total switch failure. Loss of the SU does not necessarily imply complete switch failure as long as it is fail safe to the effects of its failure.

Translation of the Heuristic Algorithm into a set of Rules

The Primary Master should be located at a point that is the shortest number of links away from the nodes on the periphery of the network.

* Rule 1: The Primary Master should be located in the 'centre' of a network, where the 'centre' is a function of distance, traffic, link and node technologies.

Taking the distance from a node to its furthest node on the network the maximum route budget is the cost of the longest routeing.

* Rule 2: Centre of a network occurs when the maximum calculated route budget is at a minimum.

The cost of a routeing comprises the total 'cost' of the nodes and links that comprise the routeing.

* Rule 3: The maximum route budget is the sum of the link and node costs taking the maximum cost between two nodes in a network.

The cost of the links is dependant upon their traffic capacity (the greater the capacity the greater the cost of failure), technology (different transmission media have

different susceptibilities to heat) and distance (the longer the link the greater the probability of its failure due to human intervention).

* Rule 4: The link budget comprises the sum of traffic capacity, wander resistance and distance costs.

The cost of a node is a function of the number of digital links connected and hence the number of link cards installed.

* Rule 5: The cost of a node is a function of the number of digital routes.

Using the Rules 1-5 it is possible to identify one or more nodes that are network centres. If only one is found, it is necessarily the centre, otherwise the choice between them is determined by the greatest number of digital lines and account is taken of its level in the hierarchy. If more than one node proves to be a 'best' location for the primary master then it is necessary to examine more closely the attributes of the nodes. For example the greater the number of digital links connected the larger the number of nodes parented, and hence directly controlled, on the site. A further test is the level in the hierarchy the nodes are, the higher the better; and, furthermore the cumulative distance of the parented nodes will effect reliability.

* Rule 6: When two or more nodes have the same lowest maximum route cost the node with the greater number of digital routes is at the 'centre' of the network.

* Rule 7: When two or more nodes have the same lowest maximum route budget and the same number of digital routes, then the node that is higher in the hierarchy is the 'centre' of the network.

* Rule 8: When Rule 7 has failed to select the 'centre', the sum of the distances of every connected link to each node is calculated. The network 'centre' is now the node having the minimum distance cost.

If there are still a number of nodes deemed suitable for the primary master, then the technology is examined and that with the best in terms of wander resistance is selected.

* Rule 9: When Rule 8 fails to select a 'centre', the technology of the links is considered and the 'centre' taken as the node having the minimum wander cost on the links.

* Rule 10: If, in Rule 9 there are a mixture of technologies in a single route, the worst case is taken as representing the whole distance.

By this stage, the Primary Master location will have been found. It is then necessary to discovered the Secondary and Tertiary Masters by repeating the above process. However, at this stage it is not necessary to find the optimum 'centre' but rather the next best two node locations that are closest to the Primary.

* Rule 11: Identify the best two neighbour nodes to the primary master node and search each, re-using Rules 1-10 replacing the master node with a cost value by zero.

The interconnection of nodes must be made to a set of rules based on the hierarchy. In addition, a number of resilience requirements must be satisfied.

* Rule 12: A node can only give a synchronisation to another node at the same level or lower level in the hierarchy.

It is also necessary to prevent looping thus the direction of control is important. Under normal operating conditions the Primary must have precedence over the Secondary and it over the Tertiary.

* Rule 13: Control between Masters will be unilateral from the Primary to the Secondary to the Tertiary.

* Rule 14: Control between nodes at differing levels in the hierarchy is unilateral from the higher to the lower.

* Rule 15: Control between nodes at the same level is bilateral.

The requirement for resilience dictates the following link requirements.

* Rule 16: Every level 3 node should be connected to a level 4 node by no less than two synchronisation links in tandem and the precedence allocated to the links is greater to those links from the higher level.

* Rule 17: If there are more than one link from the same level terminating on the node, a higher precedence is given to the link with the lowest number of links in tandem to the primary master.

* Rule 18: Level 2 nodes should have a minimum of two synchronisation links with precedence given to the link from the higher level.

* Rule 19: Level 1 nodes should receive their synchronisation from a node at level 2. If this is not possible then a minimum of two links from the same level is preferred.

* Rule 20: Where no suitable primary rate link is available, a 64 kbit/sec link to the highest level with which the node has a traffic community of interest is required.

APPENDIX 2

CASE STUDY TWO

Estimation of Demand for Telephone Service & Exchange Dimensioning and Location in the Bangkok Metropolitan Area

1 Introduction

This case study uses the three new algorithms detailed in the thesis, it also details the data contained within the demonstration program of the Plymouth Expert System.

2 Objective

The overall objective of this case study is to present reliable information on the demand for PSTN in the BMA in a form that parallels the process detailed in the main body of the thesis.

The case study includes the identification of those areas of the BMA which would experience a high growth in demand for telephone service; to qualify the telephone connection demand as existed in 1990, 1994 & 1996, and to compare this with the actual and planned capacity of TOT, (equating with their ability to satisfy that demand).

In addition, first cut designs are produced to meet this estimated demand.

3 Background

This case study is one of validating output from the Plymouth Expert System with that of the BT Tallis Consultancy, Telephone Organisation of Thailand - (TOT), the CCITT GAS 10, and the Japanese JICA study teams' results.

Various analytical tools were used by the parties including:

- 'bottom-up' demand assessment,
- opinion market research,
- international comparisons,
- socio-economic modelling, and,
- econometric modelling.

3.1 Bangkok

The BMA of 1568 square kms. is divided into 36 Districts: Bangkok's Department of Town and Country Planning has classified districts within the BMA into three groups: City Core, Suburban, and City Fringe according to their population densities and locations.

The city core consists of sixteen districts. Its population was 1.6 million in 1970 and grew to 2.3 million in 1980 and 2.4 million in 1985. Population density for the area since 1985 continues to decline as the official NSO census statistics are based entirely on residential status. On a percentage basis approximately up to 30% of the population of the area are transient or expatriate, commuting daily into the area.

Counter to population decline, telephone service demand was increasing at a much higher rate than in other areas of Bangkok.

The redevelopment of the city core area has continued unabated since its conception in early 1987 and seems likely to continue for the future. The majority of the property expansion were concrete high-rise commercial buildings, with a higher than average telephone demand.

The suburban area comprises thirteen districts. Its population has grown rapidly during the last fifteen years having increased from 1.3 million in 1970, to 2.1 million in 1980, and 2.6 million in 1985. As a result the suburban share of the total Bangkok population increased from 41% in 1970, to around 47% in mid 1985. Of the three zones within the BMA, the suburban area had the highest population growth rate averaging 4.9% per annum between the years 1970 to 1980, decreasing slightly to 4.6% during the last 5 years. Almost half of that growth was concentrated within just two districts Bangkhen and Bangkapi, with Bangkapi having the highest population growth, averaging over 10% per year between 1970 and 1985, followed by Bangkhen which has grown at an average rate of 6.7% per year; population densities in suburban district range from 1,600 to 9,800 persons per square kms., with a mean of 4,500.

4 Design Methodology

4.1 New Demand Assessment

New demand for telephone service is correlated primarily with new building developments and secondly with the more intensive use of existing buildings. A

market opinion survey was conducted to provide accurate and reliable data on both these elements of demand.

For the purposes of their study demand comprise different sources as follows:

- New Major Developments (office buildings, condominiums or housing estates). This comprises projects under construction, projects which had planning permission, assessing the likelihood of the their construction, those unplanned but likelihood of their construction, and those unplanned but likely to implemented. Telephone lines provided by building or housing estate developers. Additional lines required by occupants of those developments immediately or soon.

- Existing Major Developments. Unfulfilled demand, i.e. lines applied for but not supplied and demand for additional lines in existing businesses.

- General Demand From Smaller Developments

- Fulfilment of 'Suppressed' Demand in:

Office Buildings: First determine the total area of commercial buildings under construction and the probable mix of occupants by type of business. Thence, determine the likely requirement for telephone lines, per square metre, for different types of business and overall: multiply the above two and deduct the number of direct external lines and switch board extensions (insofar as they were considered substitutes for direct external lines) provided by the developer to calculate the suppressed demand for services.

Government Offices: Similar process to commercial buildings with a smaller number of lines required per square metres.

Residential: Assume that each house or apartment in a high rise building would require one line.

Hotels: These were a special case since they require a large number of rooms and line demand is not in the same proportion to the number of square metres or space in the building as for commercial and residential.

Private Hospitals: These were similar to hotels but with a smaller number of telephones required per room.

Shopping Centres: Assume that each shop would require one telephone and that each department store would require a number in proportion to its selling area.

Number of Lines Per Square Metre and Other Estimates: This information was obtained from a sample survey conducted amongst a mix of business types in Bangkok. In addition, sources such as architects, estate agents, developers and telephone equipment vendors added value.

Information obtained includes: number of telephone lines occupants wants under circumstances of limited availability and high prices. Number of telephone lines that occupants would require under circumstances of unlimited availability at reasonable prices. Number of telephone lines that occupants would require under circumstances of unlimited availability at high prices. Estimates of demand for each major building

or housing estate were then based on: Estimates by occupiers: Actual use per square metre per employee in 'mature' commercial areas (on the assumption that the newer commercial areas would eventually reach those levels): Comparisons with actual use in Singapore and Hong Kong per square metre, per employee, by type of company.

4.1.1 Research Methodology

Data was collected from two sources: Primary, i.e. interviews conducted with a sample of sources in Bangkok; Secondary, i.e. from published documents available from various government agencies.

Primary information was collected by personal interviews with three groups of respondents:

- City Planning "Experts",
- Building Managers,
- Telephone Vendors and Users.

Secondary information was collected from the following sources:

- Documents available in various government agencies including: Building permits issued by the Bangkok Metropolitan Administration; Population data from the National Statistical Office; maps published by the Bangkok Metropolitan Administration showing land usage in Bangkok, i.e. residential, commercial, industrial or government; A map published by the Telephone Organisation of Thailand which shows the areas served by their existing telephone exchanges in the BMA;

- Reports and studies on telephone demand previously undertaken by other groups,

- Newspaper and magazine actions announcing new real estate projects and developments,

- A Real Estate Study published by Thailand Investment and Securities Company which outlines the cost of land throughout the BMA.

For the purposes of the study, BT divided the BMA into three sections, City core, Suburban, City Fringe. Within each of these areas, a "micro" approach was adopted, i.e. individual buildings and developments were identified, both old and new, their occupants and their need for telephones.

Buildings were selected according to location, type and the number of floors (i.e. at least six). The interview procedure involved the use of two questionnaires designed as follows:

- The Building Questionnaire: This questionnaire asked for the following items of information from each of the building managers BT interviewed:

° The date the building was completed,
° The total usable office space within the building,
° The type of occupants, by business sector, and the floor area each occupies,
° The number of telephones provided per occupant and the number of telephone lines bought privately by each occupant.

The building questionnaire was designed to identify the following:

The mix of occupants in each building,

The number of telephones used both per square metre and per employee in each type of business.

- The User Questionnaire: This questionnaire requested the following data from each of the occupants of the selected buildings that BT interviewed:

° The number of employees, both actual and planned for the next two years,

° The total office space available, again both actual and planned for the next two years,

° The number of telephone lines (both direct and extension lines) available,

° The amount paid for each telephone line,

° The estimated demand for telephone lines (both direct and extension lines) in the next two years, subject to two conditions: the cost was 10% lower, or, the cost was 30% higher than the TOT rate.

° The number of extension lines that the company would prefer were direct lines,

° The source of telephone lines in use.

The user questionnaire was designed to provide:

°The estimated present and future demand (actual and suppressed) for telephones, given cost and availability constraints,

°Sample telephone usage ratios both per square metre and per employee in each business sector.

However, the above interview sample only told us how many telephone lines the existing buildings located in the BMA need. To estimate future demand for telephones in the city, it was necessary to examine the growth in construction, both in absolute terms and by type of building, since it was the most significant factor influencing future telephone demand in the metropolis.

To examine the future development of the BMA area, BT conducted interviews with the city planning "experts". In particular, all the construction projects examined consisted primarily of office complexes and condominiums, residential condominiums, hotels and shopping centres located both in new and "mature" business areas. From information provided by these sources, buildings over six floors under construction were examined for the following characteristics:

° Project title and location,

° Type of building, in terms of purpose,

° Number of floors and total area available,

° Target date of completion,

° Number of telephone lines required or planned for the building's completion.

Information obtained was then transferred onto a map of the BMA's 36 districts.

The existing TOT telephone exchange boundaries were also transferred onto the map. As a result, the following information was recorded within each one kilometre square of Bangkok:

° Land use (i.e. commercial, industrial, residential or government),

° The cost of land,

° The existing TOT telephone exchanges,

° The mass transit system, expressway and planned new roads,

° Buildings with at least six floors, both existing and scheduled for construction, by type.

From the information displayed on the map, a database for each map sector (i.e. one square kilometre) was created from which estimates of the present and future demand (at three and five year intervals) for telephones per square kilometre of the BMA were recorded.

Demand for buildings of less than six floors were assessed by using penetration factors for each type of building and occupant.

To determine a degree of confidence for the findings of each square kilometre, three areas were chosen for microscopic investigation: Ploenchit representing the city core, Don Muang representing suburban and Bang Kradi for the city fringe.

For this exercise, the areas were split into blocks of 200 metres square which were subjected to detailed analysis. The analysis consisted of examining every building, regardless of size, located within each of the squares to determine the opportunities for development or redevelopment and thus the current and future demand for telephone lines and use made of the penetration factors.

To make these findings relevant, the two hundred metre squares were "grossed up" to match the one kilometre square sections in the corresponding area as displayed in the BMA map.

To support the findings obtained from interviews with real estate developers, BT conducted interviews with three groups of individuals who were directly or indirectly involved in the construction development process in the BMA, to assess their opinions on both the trends of building construction and telephone usage within the city.

The first group consisted of architects, construction companies and contractors, real estate agents (with international experience) and telephone vendors. Their opinions on the following subjects were sought:

° The existing and potential growth areas in terms of project development,
° Telephone line use in "mature" (e.g. Silom, Ploenchit) vs. new commercial areas (e.g. Ratchadaphisek, Bangna-Trad),
° Telephone line use by companies in different sectors of the economy (i.e. services, manufacturing, commerce, government).

The second group consisted of commercial banks such as Bangkok Bank Ltd. and Siam Commercial Bank Ltd. who were known to be leading financiers of major construction projects. They were questioned on the following subjects:

° Which type of construction project do they finance and intend to continue supporting during the next five years,
° Which areas in the BMA show the highest growth potential by type of construction project, i.e. commercial and business office complexes, residential condominiums, shopping centres, hotels, etc.

The third group consisted of business associations such as the Thai Hotel Association, Association of Siamese Architects, Thai Contractors Association, Condominium

Business Association and Association of Private Hospitals. This group was asked their opinions on the following subjects:

° The expansion of the BMA, in general, and in particular within their specific areas of concern,

° The telephone requirements of their members.

The opinions obtained from the above series of interviews were used to supplement the findings in the database of map sectors.

To obtain some sort of relativity on the existing and potential market for telephones in the BMA, the existing telephone system in Singapore was examined for two main reasons:

- In Singapore, there was an over-supply of telephone lines, which was a complete reversal of the situation in the BMA.

- The TOT viewed the over-supply situation in Singapore as their ultimate goal for the BMA's telephone system.

4.2 Forecast of Demand

The following methods were selected as the most appropriate for use in the case study, after consideration of the availability of data and availability of CCITT case work for comparisons.

- Socio-economic modelling,
- Econometric modelling,
- Bottom up analysis, and,
- TOT & JICA Forecasts

4.3.3.1 Demand Assessment

To produce an existing demand assessment a bottom up method was used.

- The telephone areas in BMA were grouped according to districts, city core, suburban and city fringe and their land usage identified as defined by the BMA.

- For each group one telephone area was selected for microscopic investigation of housing types and usage.

- Penetration factors were then applied to estimate the telephone demand.

- The figures were then grossed up to one kilometre squares to give a telephone density per square kilometre.

- The total estimate was then refined and enhanced by:

- known building development,
- known redevelopment,
- spare land usage cost of land.

• The following exchanges were selected for detailed investigation:

- Ploenchit Exchange Area (City Core),
- Don Muang (Suburban),
- Bang Chi (City Fringe).

The estimation methodology for current, and future, demand for telephone service by square kilometre of the BMA follows.

The 36 districts of BMA were subdivided into one square kilometre blocks. Each square kilometre was defined by its general characteristics as follows:

- population,
- land value in Baht,
- land use type (primary, secondary and tertiary),
- its distance from the centre of Bangkok, and,
- demand for telephone lines in 1990, 1993 and 1995.

All estimation procedures to be discussed henceforth were adopted on the assumption that high-rise office buildings were a proxy to actual demand for telephone lines.

Demand for telephone lines in 1990 was estimated from:

- results of the survey did on the BMA's three geographical areas, namely, city core, suburban and city fringe,

- waiting list of the TOT,
- existing buildings within the BMA at least six storeys high, and,
- the number of households in each district.

From the survey results, telephone usage per square metre of office space was calculated. From the building database BT was able to determined the total usable office space in each square km. (allowing for some adjustments for those buildings below six floors.) Result of the above two were multiplied to arrive at the telephone lines in use by business and commercial buildings. Some adjustments were made to cover the residential buildings that were not condominiums as well as those buildings below six-storeys high.

Result of the above were multiplied with the number of telephone lines in TOTs waiting list to arrive at the actual demand for telephone lines in 1990.

New Major Developments in BMA required a comprehensive survey of major project developers. Information collected from interviews and desk research enabled the project team to identify those areas in BMA which were undergoing fast construction development, what types of buildings were being developed, total usable office space or floor area of each building and, to a large extent, the demand for telephone lines by each project. Estimates had been provided for incomplete data on telephone demand.

4.3.3.2 TOT Forecast

TOT provided their forecasts of connection demand, identified separately by exchange area, and a statement of exchange line units to be provided by the end of their 6th project. No year on year implementation plan was available.

The starting point for the connection forecasts was workers and waiters. The forecasting procedure as described by TOT involves reconciliation of separately initiated top-down and bottom-up procedures. TOT describe principles of field survey, penetration factors and GRP estimation.

In the provinces, historical data and comparison with similar areas were made; TOT also conduct customer interviews regularly which indicated that they would like more lines, but TOT has said that it was not unusual to find that the consumer had not even bothered to register this interest because of a recognition that lines were just not available. Organisations had also adopted a number of stratagems to overcome these inadequacies, involving, e.g. provision of their own networks, laying their own distribution cable, use of private radio systems and, particularly in recent years, use of the public cellular networks.

4.3.3.3 CCITT Gas 10 Socio-Economic Model

The CCITT GAS 10 forecasting model was derived from general socio-economic studies related to the distribution of household income and the telephone density function. Two socio-economic studies were undertaken by the CCITT in Thailand and their sample sizes used represented 0.115% of the total number of households.

In the CCITT study, use was make of surveys undertaken by the Thai NSO in 1975/76 and 1981. Due to the sample sizes used, 11514 households in 1975/76 and 11895 households in 1981, detailed investigations could only be made on a regional level. Due to the relatively limited available sample sizes, the study involved the division of the country into four different geographical areas.

For all these areas, the CCITT has distinguished between urban and rural areas, the former being represented by what is called Municipal Areas (MA) and the latter by Sanitary Districts and Villages (SDV).

Before the forecasts for the years 1995 and 2000 were made the potential demand for the years 1976 and 1981 and 1990 were calculated to validate the model.

4.3.3.4 BTs Socio-Economic Model

Data was obtained from the published Thailand Government Statistics covering Population, Gross Regional Product, Prices, Number of households and household size. Figures from the years that were available were then extrapolated to form a long-term view about the growth with time of Regional Product, Population and Prices for each region.

Although actual regional product figures could be obtained for a few years and used to forecast the future growth rates, the method was not as reliable as using a recent estimate for the growth of the economy as a whole. The regional estimates were therefore revised to take account of the estimated growth forecast and were adjusted by the difference between this figure and the initial forecast.

The GRP growth and population figures enable per capita GDP to be calculated for each region at 1990 and 1999 and this was used to estimate the expected telephone penetration. A broadly log-linear relationship was assumed to exist between the per capital GDP in US$ and telephone penetration.

Using regression analysis, the coefficients of the relationship were derived and used to forecast telephone penetrations based on current and future per capita GDP.

The figures obtained for expected telephone penetrations at 1990 and 1999 were then multiplied by the forecast regional population at those dates to give a figure for the expected demand for telephones at 1990 and 1999.

4.3.3.5 CCITT Econometric Model

The CCITT econometric model combines economic analysis with statistical data and mathematical correlation. These were combined in one quantitative process.

For their study, ultimately a model was developed using only one independent variable. Multi-collinearity relationships among the other independent variables were considered and the limited amount of data proved that it was not suitable to use more than one variable. Hence, a simplified relationship between demand and the single independent variable of GRP provided them with the overall best model.

4.3.3.6 BT Econometric Model

A series of models was developed which overcome some of the difficulties experienced by the CCITT.

A model was developed relating Residential system growth to growth in Real Consumers Expenditure and price. The model was in the form of an adjusted logistic penetration curve, assuming a saturation level of 1 connection per household.

From the available data, it was not possible to obtain a long series of strictly Business connections. The series modelled was therefore non-residential connections, which also include public call offices and service lines.

Since Business and Government lines between them comprise some 92% of non-residential connections, it was considered a suitable proxy for true Business connections in an econometric modelling context. Again a logistic penetration curve was used with a saturation of 1 connection per 0.04 Special Drawing Right (SDR) per Capita the level to which the Bank of Thailand aspires when Thailand achieves NIC status.

4.3.3.7 International Comparison

International comparisons had been made with 12 other European, African and Latin American countries selected as being comparable, from a modelling view, with Thailand in terms of GDP per capita and economic structure. The 12 countries selected compare with Thailand are as follows:

COUNTRY	GDP per Head (US$)	% Output deriving from Agriculture	Manf.	Services
Bolivia	600	23.8	34.6	41.6
Cameroon	1043	21.4	35.5	43.1
Dominican Republic	774	16.1	29.6	54.3
Ecuador	1160	15.3	36.9	47.8
Egypt	775	22.4	29.9	47.7
El Salvador	763	23.9	24.5	51.6
Guatemala	889	25.6	20.0	54.4
Honduras	740	28.4	19.2	52.4
Nigeria	800	23.8	34.0	43.2
Thailand	*771*	*22.3*	*30.0*	*47.7*
Tunisia	1140	17.2	33.5	49.3
Turkey	1120	20.5	33.9	45.6
Zimbabwe	629	13.6	34.3	52.1
Un-weighed Average (exc. Thailand)	869	21.0	30.5	48.5

The un-weighed average GDP per head of the 11 comparitor countries was at US$ 869, higher than Thailand, to reflect the strong growth prospects for the Thai economy over the next 10 years.

4.3.3.8 JICA Forecast

The basis of the residential connection forecast was the identification of households whose income levels were more than 7000 Baht a month. An estimation had to be made on the number of house holds based upon a forecast of the increasing population and the decreasing size of each household unit.

The business connection forecast was based upon the number of employees who had no less than upper secondary school education level; number of connections per employee in USA, UK, Canada, Sweden and Japan; calculation of connections on a pro-rata basis assuming one employee in the above named countries equate to one Thai employee of at least secondary level education.

JICA also developed two further 'expressed demand' models for Bangkok and Rest of Thailand based upon existing connections and waiters.

5 Bangkok

The BMA of 1568 square kms. is divided into 36 Districts. Bangkok's Department of Town and Country Planning has classified districts within the BMA into 3 groups: City Core, Suburban, and City Fringe according to their population densities and locations. They are:

LOW DENSITY RESIDENTIAL

HIGH DENSITY RESIDENTIAL

COMMERCIAL

INDUSTRIAL AND WAREHOUSE

AGRICULTURAL

The districts that comprise Bangkok are as follows:

CITY CORE 16 districts (populous central areas)

1. Yan-na-wa
2. Sathorn
3. Bang-kor-lam
4. Bang-rak
5. Patumwan
6. Pha-ya-tai
7. Rajtevee
8. Huay Kwang
9. Dusit
10. Bang-sue
11. Phra-na-korn
12. Pom-prab
13. Sumpuntawong
14. Bangkok-Yai
15. Thonburi
16. Klong-sarn

SUBURBAN 13 districts (lower population density than city core)

1. Bang-Khen
2. Don-Muang
3. Ja-tu-jak
4. Bangkapi
5. Lad-prao
6. Bung-koom
7. Rat-boo-ra-na
8. Phra-ka-Nong
9. Klong-Toey
10. Pra-vet
11. Pa-see-jaroen
12. Bangkok-Noi
13. Bang-plud

CITY FRINGE 7 districts (agricultural area)

1. Nong-jok
2. Meen-buri
3. Lad-Kra-bung
4. Ta-ling-shun
5. Nong-kam
6. Bang-Khun-Tien
7. Jom-tong

TOT has defined the Bangkok Charge Group as the BMA and the provinces of Patumthani, Nonthaburi and Samutprakarn. CAT has defined its Bangkok Charge Group as postal areas covering the BMA and Samutprakarn. MOC has defined a Greater Bangkok Area as the BMA and the five adjacent provinces: Nonthaburi, Samutprakarn, Patumthani, Samutprakarn and Nakornpathom that is equivalent to the BMA.

5.1.1 City Core

The city core consists of 16 districts. Its population was 1.6 million in 1970 and grew to 2.3 million in 1980 and 2.4 million in 1985. Since that time there has been a decreasing trend in the city cores' share of the total BMA population.

The classification of the districts directly related to the city core area are those of Saturated Urban Area or Slow Growing Urban Area.

Population density for the area continues to decline as the official NSO census statistics were based entirely on residential status. On a percentage basis approximately up to 30% of the population of the area were transient or expatriate, commuting daily into the area.

The redevelopment of the city core area has continued unabated since its conception in early 1987 and seems likely to continue for the future. The majority of the property expansion was in concrete high-rise commercial buildings, with a higher than average telephone demand.

On this basis, the traditional concepts of long term planning for telephone demand, which can be successfully implemented in areas of regulated growth, proved to be totally inaccurate in this high demand commercial environment. To achieve a reasonable situation of service on demand, TOT planned for a much higher than normal percentage of installed spare plant than the forecast predicted.

5.1.2 Suburban

The suburban area comprises thirteen districts, its population has grown rapidly during the last fifteen years having increased from 1.3 million in 1970, to 2.1 million in 1980, and 2.6 million in 1985. As a result the suburban share of the total Bangkok population increased from 41% in 1970 to around 47% in mid 1985. Of the three zones within the BMA, the suburban area had the highest population growth rate averaging 4.9% per annum between the years 1970 to 1980, decreasing slightly to 4.6% during the last five years. Almost half of that growth was concentrated within just two districts Bangkhen and Bangkapi, with Bangkapi having the highest population growth, averaging over 10% per year between 1970 and 1985, followed by Bangkhen which has grown at an average rate of 6.7% per year.

The classification of the districts is as follows:

- Rural Areas (Nong Choke, Minburi, Lat Krabang).
- Transitional Areas (Bang Khun Thian, Taling Chan, Nong Khaem).
- Fast Growing Urban Areas (Yan Nava, Huai Khwang, Phra Khanong, Bang Khen, Bang Kapi, Phasi Charoen, Rat Burana).
- Slow Growing Urban Areas (Bang Rak, Dusit, Phaya Thai, Thon Buri, Klong San, Bangkok Noi, and Bangkok Yai).

- Saturated Urban Areas (Phra Nakhon, Pom Prap, Sattru Phai, Pathum Wan, Samphanthawong).

Population densities in suburban district range from 1,600 to 9,800 persons per square km. with a mean of 4,500.

The districts classified as fast growing and transitional were projected to grow the fastest, at a rate over 3% annually during the sixth plan. Fast growing urban areas and transitional areas share of the total BMA population was projected to increase from 46% in 1980, to 53% by the end of the sixth plan, whilst the proportion accommodated in the saturated and slow growing areas was projected to decline from 51% to 44% during the same period.

The bulk of new growth was projected to occur in fast growing and transitional areas. Between 1980 and 2001 their population was expected to increase to some 2.2 million thus accounting for some 14% of the total BMA growth. In particular the population of the combined area of Huai Khwang, Phra Khanong, Bang Khen and Bang Kapi district, to the east and the north east of Bangkok was projected to grow by 1.4 million, i.e. 14% of the total BMA growth during 1980 to 2001. The latter two were areas where major investments in urban transport had taken place or were recommended for the sixth plan period. It was expected that the population density of the high growth amphoes would double on average between 1980 and 2000 with Bang Khen adding more population than any other amphoe over the period.

The population of Pathum Thane was 0.33 million in 1980 and was projected to increase to 0.68 million in 2001. The population growth was expected to be distributed relatively evenly among the amphoes. Some 30% of the total changwat

increase between 1980 and 2001 was expected to be in Klong Luang followed by Thanyaburi and Amphoe Muang which were projected to have 25% and 14% of the total population increase. It was forecast that Thanyaburi and Klong Luang would have a growth rate averaging 4% per year. The growth rates of Lam Luk Ka and Amphoe Muang were also expected to be high and to continue to increase over the period between 1986 and 2001 in part a reflection of the location of infrastructures and industrial zones projected for Pathum Thane.

5.1.3 City Fringe

The city fringe districts were potentially one of the greatest population growth regions in Bangkok. Concerning the National Economic and Social Development Board 6th National Plan; Changwat Pathum Thane, for example, was expected to increase in population at a rate of 3.52% per annum over the period 1986 - 2001, with 25% of the total increase expected within amphoe Muang. The rate of growth takes no account of the expected continuation in the rapid industrialisation of the Bang Kradi area.

The trends identified by Cables De Lyon field studies indicate an average growth in the demand for telephone service in the area around investment the, of 14.4% per year from 1986-1991 and then 9.9% until the year 2001.

5.1 Bangkok at the Micro Level

THE PLOENCHIT EXCHANGE AREA

Ploenchit Exchange is situated in the city core of BMA. It covers an area of 15.4 square kms. It is situated in what is considered one of the most prestigious parts of

Bangkok, and can be classified as almost completely business. The three districts: Phayathai, Patumwan and Phrakanong combined within the exchange area as part of the city core is classified by the BMA as saturated or slow growing, although accurately applied up to 1987, the area has since seen a considerable amount of property development resulting in a very high un-forecasted telephone demand.

The majority of the areas population were transient, either tourists resident in the many 'high' and 'medium' class hotels, or workers, employed in offices, banks or department stores.

The area also encompasses many of the foreign embassies located in Bangkok.

Over the last two years many of the older residential properties with large gardens had been demolished, and redeveloped with high class residential condominiums, business enterprises, or hotels: examples of this type of redevelopment were the World Trade Centre and the rebuilding of the Erawan Hyatt hotel.

The BMA expect that the changes would continue for the future with on ongoing trend in the conversion of existing properties into high class, multi storey buildings encompassing residential, business or retail units.

Ploenchit Exchange area has achieved more than average growth since the implementation of the 5th NESDP plan.

Because of the continuing development trend, it has been necessary for TOT to monitor constantly the area to satisfy demand as required by customer request, or by TOT Planning Staff. With their familiarity with the exchange area, it has been

possible for TOT to establish, with a reasonable degree of accuracy the growth of the telephone network.

Development of the exchange area can be divided into 2 classifications:

- The traditional growth in demand which was derived from an increase in population, better working conditions, and higher living standards.

- The new business demand which was based on a forecasted requirement, derived from the projected utilisation of new buildings.

The first classification applies in the main, to existing residential property and small businesses, which were served from distribution cabinets. The growth trend in this instance was estimated at 2% overall, but actual total numbers were declining as more and more property was demolished for the construction of new buildings.

The second classification demand was established by analysing request to TOT commercial section by customers, i.e. a stated telephone requirement. The demand is served either by the provision of large cables terminated on a remote digital frame, or if the projected demand was expected to be of a very high order, by the provision of a remote switching unit.

Under the development programme, Remote Switching Unit (RSU) areas had been allocated within the exchange area:

- Indra Regent Complex
- Hollywood Shopping Complex

- Krung Thai Bank Head Office
- The World Trade Centre

Where demand was required to a single unit construction, but was of lessor order than that required for the provision of an RSU, service was generally provided by a single direct feed cable.

Speculative development was included in the TOT forecast to allow for telephone demand arising from unknown developments which would be constructed during the 10 year planning period.

With many open land spaces still available development can be expected anywhere. However, with the probability that all the known developments had been included no speculative development was accounted for between Base and Base + 5 in the forecast of demand. TOT assume a slow down in developments, one average size (100 Units) building was speculated each year between Base + 5 and Base + 10.

For providing the high-rise buildings with 'goods and service' some speculative development was forecasted.

A complete top down and bottom up the assessment of demand for the Exchange area of Ploenchit was carried out by TOT consultants D&N.

They obtained an updated waiting list from TOT Commercial section and all applications for service investigated and identified.

Applications for high demands were considered according to the application concerned, other property units which were categorised as requiring investigation, were visited and the responsible person located and requirements discussed.

These investigations indicated that the actual demand at Base and the projected demand at Base + 5 produced higher penetration factors (PF) than those used by TOT. On this basis, adjusted figures were used; a summary of the PFs and demand densities are given at the end of the case study.

The general terms the following situation exists in Ploenchit exchange area.

Total customers connected	(approx.)	52,500
Total waiters	(approx.)	17,000

DON MUANG EXCHANGE AREA

Don Muang exchange area is situated to the suburbs of BMA in the district of Bangkhen. It covers an area of 12.1 square kms.

Within the exchange boundaries are situated the Bangkok International Airport complex, The Royal Thai airbase of Don Muang and an integrated Remote Switching unit known as Muang EK. The area, designated as a fast growing suburban, is a mixture of existing residential housing of medium and high classification. Similar projects were under construction in area's allocated for development.

The industries tend to be of medium size catering for domestic needs and the local consumer infrastructure which is always present in high density populated areas. These were located in three storey shop houses bordering the main roads.

The combined land usage of the projects totals 5,300 rai and on completion would have added 16,886 medium and high class building units to the area. In addition, areas which were considered as existing residential, which over the last few years had experienced rapid expansion, but were not yet fully developed, were expected to be completely in-built within the five year planning period.

With the exception of Muang EK. which was not included in the 42 projects, although a large extension to the area was anticipated, most of the future expansion would be to the east of Phaholyothin road or to the north of Lam Luk Ka road.

Further information, regarding future trends for the area, was obtained from the office of Town and Country Planning and the National Housing Authority.

The National Housing Authority the government body responsible for the housing of the homeless and lower income families, has identified areas where housing commiserate to the needs of the community and their ability to support themselves can be constructed near to support facilities, and good communications. In the 6th Development Plan about 50,000 housing units would be required each year in 14 areas suitable for this purpose, three were situated in Bangkhen.

Due to the geographical complexity of the exchange area and the demand requirements, it has been impossible to consider the exchange area of Don Muang as

a separate entity within its defined boundary. Areas exist within this boundary to which service cannot be readily provided within the normal transmission limits.

The existing site locations for the exchanges of Thanyaburi and Lam Luk Ka, were such that the geographical separation is 16.5 kms. and 11 kms. respectively.

Ratanat Kosin, is an existing development situated on the Pathum Thane road north of the Don Muang Exchange area, within the existing defined Rangsit exchange boundary. At the time there were recorded 2,885 existing building units. Space exists within the estate layout for property expansion which would certainly be carried out before Base + 5. The distance to the site entrance from Rangsit exchange is 5.6 kms. Rangsit has an existing 1,918 customer base.

The penetration factors (PF) were calculated from a statistical study of various areas within the existing exchange area of Don Muang. The areas selected for analysis comprised existing developed properties, with little or no waiting list, and where telephone service was not subjected to a lack of line plant. In total 71 areas of mixed housing units were identified, and the existing service situation investigated. From these results the average PF for the exchange area were calculated.

The figures indicate an average service growth of 7.8% which was consistent with the area development concept.

This data was combined with other general survey of other Suburban Areas to give the overall penetration factors.

BANG KRADI EXCHANGE AREA

Bang Kradi exchange is located in the administrative district of Pathum Thane, in the city fringe of Bangkok. Concerning the 6th National Plan, Changwat Pathum Thane was expected to increase in population at a rate of 3.52% per annum over the period 1986 - 2001, with 25% of the total increase expected within amphoe Muang Pathum Thane district, which itself was expected to increase at the slightly lower growth rate of 2.82%

This rate of growth which was one of the highest within the metropolitan area, was attributed to the expected continuation in the rapid industrialisation of the Bang Kradi area.

To gain an insight into the rate of growth of future building, it was first necessary to assess the increase in population by application of acceptable growth factors. All such factors used in the following projections were taken directly from the following sources.

'Population and Housing Census' by the Faculty of Political Science at Chulalongkorn University, in conjunction with the National Statistics Office for the 6th National Plan.

- Population growth = 2.82% per year.
- People per household. = 5.8 reducing to 5.2 by 2000
- Residential houses at time of survey = 2099

To establish demand for a 10 year provision period, it was also necessary to take into consideration vacant lots (in-fill sites) and empty areas of land (speculative sites) that were expected to develop within the forecast period. In-fill sites, 485 at Base + 5, 566 at Base + 10 and speculative sites, 1965 at Base + 10

As it was the propose of the Forecast to evaluate demand for a provision period of 10 years and to minimise the potential numeric recalculation of cabinet areas, part of the empty areas of land within these cabinets has been evaluated as a speculative site, be it Speculative Residential (SR) or Speculative Business (SB).

However, should the calculated demand for any cabinet area be formed mainly from these SR/SB sites, they can be deferred until the actual development of these sites occurs and creates demand sufficient to warrant the installation of the cabinet. By calculating cabinet areas in this manner, a realistic 10 years provision period can be achieved and as these sites were only evaluated at Base + 10, they do not in any way in influence the initial primary cable instalment.

During the detailed "Field Survey", it became evident to the BT team that Bang Kradi is developing at a fairly rapid rate. As previously mentioned, this was probably due to the increasing number of industrial facilities that had already been established or were either under construction or planned for the future.

The industrialisation has generated many new employment opportunities which had led to a growth in housing development and commerce.

6 Demand and Forecast Study Results

6.1 Market Opinion Survey

Information obtained from existing documents include:

- Population data in each of the 24 districts of the BMA from the National Statistical Office.

- List of projects given building permits around the BMA during the period January 1980 to January 1990 from the Building Regulations Division of the Bangkok Metropolitan Administration.

- List of construction projects awaiting approval from the Bangkok Metropolitan Administration;

- Housing projects under the sponsorship of the National Housing Authority from 1987-1991.

- New roads to be constructed and old roads to be expanded from the Bangkok Metropolitan Administration,

- Bridges to be built from the Department of Public Works of the Ministry of Interior,

- Route plan of the second stage expressway from the Express and Rapid Transit Authority of Thailand,

- Land use maps from the Bangkok Metropolitan Administration,

- Telephone exchange maps from the Telephone Organisation of Thailand, and

- Cost of land around the BMA from a report compiled by the Thailand Investment and Securities Co., Ltd.

Interviews conducted were of two types:

- Type A refers to interviews with government planners, project developers, business associations and other organisations such as commercial banks, architects, telephone equipment vendors who were directly or indirectly involved in the development process of the BMA, and,

- Type B refers to interviews with users of telephone exchanges including building managers or owners and occupants of the building.

- Government Planners
- National Housing Authority
- Public Works Division, Bangkok Metropolitan Administration
- Department of Public Works, Ministry of Interior
- The Expressway and Rapid Transit Authority of Thailand
- Project Developers
- Associations
- Architects and Contractors
- Equipment Suppliers

- Commercial Bankers
- Real Estate Agents

6.2 CCITT GAS 10 Socio-Economic Model

The CCITT GAS 10 forecasting model was derived from general socio-economic studies related to the distribution of household income and the telephone density function. Two socio-economic studies were undertaken by the CCITT in Thailand and their sample sizes used represented 0.115% of the total number of households.

In the CCITT study, use was make of surveys undertaken by the Thai NSO in 1975/76 and in 1981 and covered the whole country. Due to the sample sizes used, 11,514 households in 1975/76 and 11,895 households in 1981, detailed investigations could only be made on a regional level.

Their study of household expenditure distribution and telephone density functions were made for the following areas:

Bangkok Metropolis

Northern region

North-eastern region

Central region

Southern region

For all these areas, the CCITT has distinguished between urban and rural areas, the former being represented by what is called Municipal Areas (MA) and the latter by Sanitary Districts and Villages (SDV): Before the forecasts for the years 1995 and

2000 were made the potential demand for the years 1976 and 1981 and 1990 were calculated to validate the model.

6.2.1 Basic Data Required

The data needed for the forecasts were taken from several sources and related to the long-term plans drawn up by the Government on the social and economic development of the country.

The following information was used in the study:

- Development of population and households

The potential users of telephone services for the residential use were the households. The long-term evolution of population and household size was studied to forecast the number of households by regional distribution, social character, professional category or any other division desired (and available from the sample survey data sources).

- The household expenditure (or income) distribution

From the socio-economic surveys mentioned previously, the frequency distribution of expenditure (income) of the households was known. These functions were derived for regions of the country, urban or rural areas, ethnic groupings, professional categories, etc. as necessary for the investigations.

The long-term trend of expenditure (income) was investigated by utilising estimated from the government plans on consumption and investment development. The

development trends on consumer price index (CPI), also a useful measure for this development trend on household expenditure.

• Telephone service density function

The density function was calculated from the above-mentioned socio-economic sample surveys. It shows how the use of telephone services was related to the expenditure (income) level of the households. The same kind of stratification's used for the household expenditure distribution were also used.

6.2.3 Forecast of Potential Demand

The base document for CCITT's forecast of population development in Thailand was the ESCAP publication, Country Monograph Series No. 3, Population of Thailand. As the document was relatively old, 1976, CCITT complemented it with more recent data, including the Fifth National Economic and Social Development Plan (1982-1986).

Since it was of particular interest to know the development of households by region and subdivided for urban (MA) and rural (SDV) areas, the CCITT studied available data from 1960, 1970, 1976, 1980/81 and 1985 to forecast the demand of telephone services for the residential users.

A factor of importance was the Thai Government plan to reduce the population increase to 1.5% after 1986. CCITT has considered a reduction of the population increase that approaches this value.

As the use of telephones, is directly related to the number of households, it was important to note that the size of the households was gradually decreasing. CCITT considered that 5.2 persons per household would gradually decrease to 4.5 persons per household in the year 2000. This figure was for the country total which was tentatively used for the regions, although it was realised that the size of the household would vary for regions and for urban or rural areas. CCITT, felt that the differences were not important.

6.2.4 Residential-Business Telephone Users

The CCITT surveys used for investigating telephone demand were based on interviews of households. It covers therefore mainly the residential type of users.

The business and official telephone users were therefore not included and had to be estimated separately. Since this user category is usually the dominating one in countries with a relatively low telephone penetration great care has to be taken to evaluate the future development.

When the demand of telephones can be provided to both the main categories, residential and business, in an unbiased way, i.e. business users were not given priority to obtain telephones nor were the installation charges such that they in an indirect way create a biased satisfaction of demand: during such circumstances the trend function of the penetration of residential users (of the total number of telephone users) follows a curve similar to the one shown in figure 81.

The same kind of function for the case of Thailand is shown in figure 82 for the cases of Bangkok and the Provinces. This curve reveals that the relative number of

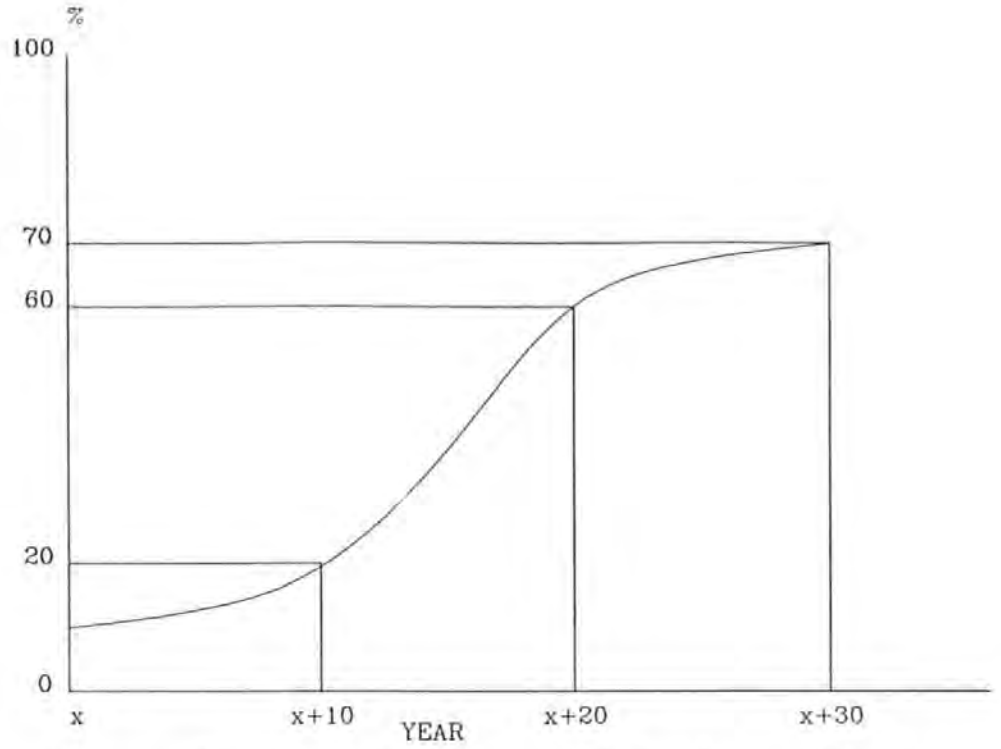


Figure 81 Penetration of Telephone Service (Japan)

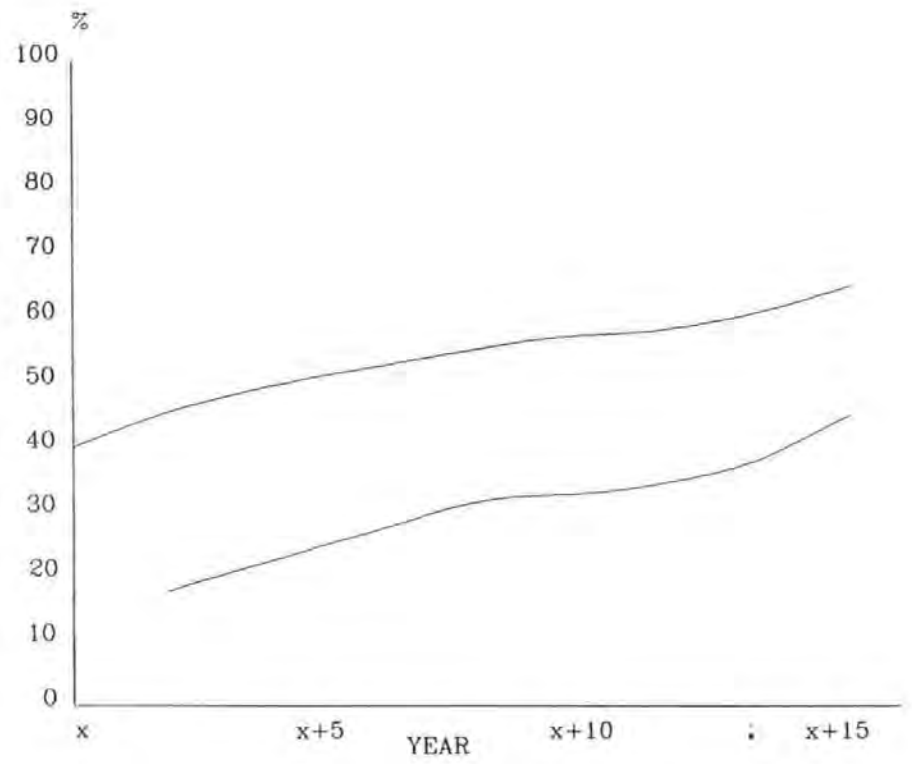


Figure 82 Penetration of Telephone Service Bangkok and Provinces

residential telephone users was higher in Bangkok than in the Provinces. It was also clear that until 1982 the business and residential telephones in Bangkok were almost stable at 50/50, whilst in the Provinces the residential telephones had continued to progress.

6.2.5 Household Expenditure Frequency Functions

The demand for telephone services is closely related to the expenditure (or income) of the household. To forecast the demand, one has to know both the household expenditure function and the telephone density function.

From the socio-economic surveys, CCITT obtains the expenditure functions for the four regions and BMA.

These functions had been calculated for the situations at years 1975/76 and 1981, corresponding to the two survey years.

The density functions were relatively stable in time and were assumed to have the function as given in the tables for 1981.

The main reason for the increase of demand, at similar tariff situations, lies in the fact that the expenditure (and income) increases by time. This increase is mainly due to inflationary reasons. The data for 1975/76 and 1981 follow very closely the CPI index changes, e.g. the Median point of the accumulated household expenditure function was 1855 Baht for the Kingdom in 1975/76; it moved to 3199 Baht in 1981. This corresponds to a factor 1.725 (3199/1855) which corresponds very well to the CPI change from 1976 (= 100) to 1981 (=172.1).

6.2.6 Service Density Functions

Investigations by CCITT in Thailand and three other countries in Asia had shown that the expenditure functions and density function were both simple expressions of a logistic type and well explain actual data.

As mentioned earlier, the density functions tend to remain stable by time, contrary to the expenditure functions.

It has also been found that density functions in different countries were found to be similar, expressed in the same currency. The example from Bangkok and Singapore shows how the data from telephone density and TV density were very similar in both cases.

The calculated parameters for the expenditure and telephone density functions, for the various regions, as well as the calculated average percentage demand, were given in detail for the years 1975/76 and 1981.

6.2.7 Demand PF Telephones in 1975/76 and 1981

Before calculating the forecasts for 1990-2000, the CCITT investigated the demand of residential and business telephone users for these two years. Since the demand of these was considerably higher than the actual statistical figures on working lines show, it was important to have this evaluation as starting points for the forecasts for the next 15 years.

Inflation usually makes the income function (expenditure function) grows rapidly by time and that usually the telephone tariffs increase slower, makes the tariffs appear lower relative to income. This has a catalysing effect on demand of services from the lower income groups of households. One can therefore expect that the potential demand in the future would grow fast. How well this demand can be satisfied is a question of financial means available and Government policy towards the telecommunication sector. The Thailand Government places high priority on the development of the telecommunication services.

The demand for telephone services in 1975/76 and 1981 has been evaluated from the available statistical data and the household expenditure surveys. The figures were considerably higher than the actual working customer lines in those years. The fact the major part of the population of Thailand lives in the rural areas of the country makes these areas the greatest sources of new telephone demand in the future.

6.2.8 Forecasts for 1985, 1990, 1995 and 2000

From the detailed data for years 1975/76 and 1981, the CCITT was able to define the expenditure and density functions for these years. The potential demand for residential and business users was also calculated. They were able to evaluate the potential demand in those years which formed a starting point for their forecasts for the next 15 years.

As the density functions were relatively stable over time, they applied for the forecasting period, the same density parameters as obtained for 1981. From the 1975/76 and 1981 data, one sees sometimes a variation in the B constant. If the B constant increases (in absolute value), the A constant also increases such that factor

A/B remains almost the same. This means that the median value for the density function is unchanged (Median = exp(-A/B)).

The change of the expenditure function is thus the most important function to estimate for the future years.

For the year 1985, the official CPI value (= 194.4) which was used to evaluate the change in expenditure per household for 1985.

Their 1990 value for CPI was calculated from the statistical data on CPI for the period 1976 to 1985, using the double exponential smoothing method, being 222.7; which was accurate as the actual value for 1989 was 221.7.

For the other five year periods, they assumed that the expenditure per household would increase yearly by 4% (1990-1995) and 5% (1995-2000).

The development of a relationship between residential and business users was another important and difficult, matter. That this relationship had not reached equilibrium obliged the CCITT to hypothesise on the evolution. As they were forecasting the potential demand, they assumed that the business users reach about 20% of the total for the country totals. The business users were further considered to be relatively larger in the urban areas than in the rural areas. Also the particular case of Bangkok, with a higher business proportion has been considered, as the available statistical data also show.

6.3 BTs Socio-Economic Model

Data was obtained from the published Thailand Government Statistics for each region in question under the headings Population, GRP, CPI, Number of households and household size. Figures from the years which were available were then extrapolated linearly to form a long-term view about the growth with time for each region.

Although actual regional product figures could be obtained for a few years and used to forecast the future growth rates, this method was not as reliable as using a recent estimate for the growth of the economy as a whole. The regional estimates were therefore revised to take account of the estimated growth forecast and were adjusted by the difference between this figure and the initial forecast.

The GRP growth and population figures enable per capita GDP to be calculated for each region at 1990 and 1999 and this was used to estimate the expected telephone penetration. The broadly linear relationship between the per capital GDP in US$ and telephone penetration was assumed noting the reservations of this method by the CCITT.

Using a linear regression analysis, the coefficients of the linear relationship were derived and used to forecast telephone penetrations based on present and future per capita GDP.

The figures which were obtained for expected telephone penetration at 1990 and 1999 were then multiplied by the forecast regional population at those dates to give a figure for the expected demand for telephones at 1990 and 1999. These figures were

also used to derive the annual growth which must be achieved to increase the number of connections from that at 1990 to that given at 1999.

The figures show that by far the greatest potential demand was in metropolitan Bangkok and this was born out by its central position in Thailand's economy. Of the other regions examined, the central region has the smallest number of connections, because of its relatively small population, but has the highest growth rate outside metropolitan Bangkok due to buoyant economic growth and a population growth rate which was higher than average. Conversely, the North East has the largest number of telephones because of its large population, but has the slowest growth rate due to relatively sluggish economic performance.

6.4 CCITT Econometric Model

The econometric model combines economic analysis with statistical data and mathematical correlation. These were combined in one quantitative process.

The general model as recommended by CCITT can be expressed mathematically as follows:

$$Y_t = a_t + b_1 x_{1t} + b_2 x_{2t} + ...+b_n x_{nt} + e_t$$

where

Y_t is the dependent variable, in this case (demand),

a_t is the constant term (intercept) in the equation,

$b_1, b_2...b_n$ were coefficients expressing the quantitative relationship between Y_t and X_{nt}

e_t is a term to account for error,

x_{1t} , x_{2t}..x were independent variables,
typical examples include gross domestic product,
consumer capital investment expenditures,
population, number of households, etc.

The model requires a calculation of the intercept and coefficient of the independent variables. The calculation can be performed by the techniques of multiple regression.

The following steps illustrate CCITTs application of its econometric model in forecasting the future demand for provincial telephone service in Thailand.

- The availability of data to CCITT was very limited. The economic independent variables considered included population, number of households, gross domestic product and gross provincial product. Other potentially important factors such as capital investment and consumer expenditure were not considered since reliable forecasts of these variables did not exist.

- A hypothesis was formulated to relate the independent variables x_{1t} , x_{2t} , x_{3t} , x_{4t} to the dependent variable Y_t.

For their study, ultimately a model was developed using only one independent variable. Multi-collinearity relationships among the other independent variables considered and the limited amount of data proved that it was not suitable to use more than one variable. Hence, a simplified relationship between demand and the single independent variable of 'gross provincial product' provided them with the overall best model.

The final form of the model was:

$$Y_t = aa\ x^{B_t}$$

which was derived from the linear relationship

$$\text{In } Y_t = aa + B \text{ In } xt$$

where Y = Total customer demand in year t,

a = Constant terms,

aa = In a,

B = Coefficient,

x = Independent variable (e.g. gross provincial product) in year t.

- The method of ordinary least squares was used to estimate the parameters a and B of the regression equation. The coefficients of the model wee estimated from historical values of demand and GRP. As historical data was very difficult to obtain; a pooling of cross sectional and time series data was introduced. This pooling increased the reliability of the parameter estimates by increasing the degree of freedom.

In the analysis, pooling uses the observations for some number, N, of different primary centre (PC) areas over a number (T) of time periods. For the analysis several primary centres were grouped together, based on common characteristics. In their study, 15 such groups were considered.

As a result, the following constant term and coefficient were estimated using the ordinary least squares method:

$$\text{In } Y = -2.929 + 1.429 \text{ In GRP}$$

The results for the forecasted demand using this econometric model for a sample PC is shown in figure 83. The graphical results of the forecast for GRP, as obtained from external sources, were also illustrated in figure 83.

6.5 The BT Econometric Model

A second series of models based upon the BTs UK model were developed which overcome some of the difficulties experienced by the CCITT. It used as base data the following economic forecast as input data.

It assumed that economic prospects and performance in Thailand were excellent. In the next decade Thailand is likely join the ranks of the Newly Industrialised Countries. However, Thailand is both commodity and export dependent. In 1988 exports made up 29% of GDP, and 25% of merchandise exports were commodities (Rice, Rubber, Corn, Tapioca products, Sugar and Tin). The forecasts had been made on the basis of 3 'scenarios' concerning the Thailand economy They were:

- An optimistic case, based on world trade growth of 7-8% per annum.

- A central case, consistent with BT's economic assumptions for its Corporate Strategy Review. World trade grows by 5-6% per annum.

- A pessimistic case, based on world trade growth of 4-5% per annum.

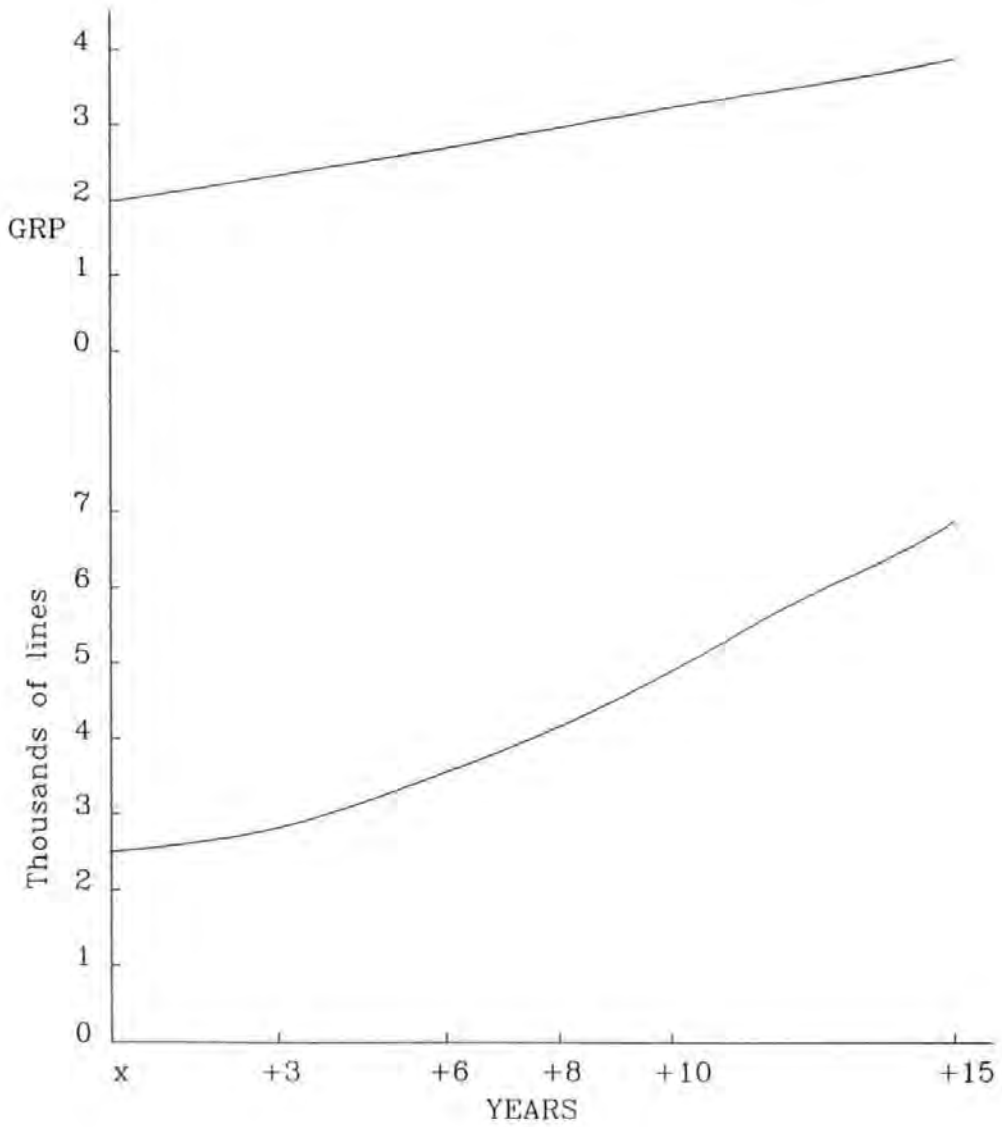


Figure 83 Forecasted GRP & Customer Demand using econometric model

6.5.1 Residential Connections Growth

Residential telephone penetration, defined as connections per 100 households, over the period since 1968, has grown steadily, and by 1987 had reached 6% of households. This growth, however, masks the inequality of telephone penetration in Bangkok (estimated at 36%), and the provinces (estimated at 2%). Central case demand in 2000 has been assessed at 1 line per households outside (based on growth to date), to yield an overall penetration of 17%.

As in other studies, it has been assumed that market development can be characterised by an S shaped curve reflecting:

- An externality effect, as more households obtain telephone service, the product becomes more visible, and of more value to a potential customer, as more people were contactable

- A finite market effect, most households were unlikely to need more than 1 telephone line and, as reported above, a more realistic saturation might be 15%.

A model has been developed relating Residential system growth to growth in Real Consumers Expenditure and price. As in other studies, the model was in the form of an adjusted logistic penetration curve, assuming a saturation level of 1 connection per household.

6.6 CCITT Logistic Model

In forecasting telephone penetration in a developed area where demand was expected to saturate in the future, such as the metropolitan telecommunication area of the country, a logistic curve model was considered the most appropriate. The logistic model accounts for a future situation where there would be demand saturation, and hence, a gradual decrease of the annual growth rate.

The logistic model can be expressed mathematically as follows:

$$\frac{Y}{N} = \frac{k}{1 + me^{-at}}$$

where Y = Total customer demand in year T,

N = Total population in Year T,

T = Forecast time unit,

m,a = constants,

k = Limit value.

The value of k is directly related to the number of both residential and business telephone users. For a metropolitan telecommunication area, the k -value should be larger than the k value used for national demand forecasting, because the metropolitan telephone density is generally higher than the national average.

For Bangkok metropolitan telecommunication, the future number of persons per household was estimated to be three, (in 1990 it equalled five). Furthermore, it was estimated that at a saturation there would be 1 telephone per household. With these

values, the value of k for residential users equals 1 telephone per 3 persons per household, .or. 0.33. The value of k for business users was assumed to be approximately 1/2 of the residential k value, 0.17. Hence the composite k value for both residential and business users is as follows:

$$k = k \text{ residential} + k \text{ business} = 0.33 + 0.17 = 0.50$$

This value for k allows the logistic curve model to weigh adequately the impact of business customer demand which is an essential consideration for a metropolitan telecommunication area with the economic characteristics as found in up country. The generalised customer demand curve, using the logistic model has three phases of growth, i.e., starting, rapid, and saturation.

6.7 BT International Comparisons

The analysis has been based on statistics of telecommunications demand and revenue obtained from TOT Report and Accounts, and statistical report, supplemented by ITU Yearbooks of Common Carrier Telecommunications Statistics. Economic Indicators were obtained from IMF Financial Statistics.

6.7.1 Basis for International Comparison

The International comparisons had been made with 12 other European, African and Latin American countries selected by BT as being comparable with Thailand in terms of GDP per capita and Economic structure. The 12 countries selected compare with Thailand as shown in below.

COUNTRY	GDP per Head (US$)	% Output deriving from Agriculture	Manf.	Services
Bolivia	600	23.8	34.6	41.6
Cameroon	1043	21.4	35.5	43.1
Dominican Republic	774	16.1	29.6	54.3
Ecuador	1160	15.3	36.9	47.8
Egypt	775	22.4	29.9	47.7
El Salvador	763	23.9	24.5	51.6
Guatemala	889	25.6	20.0	54.4
Honduras	740	28.4	19.2	52.4
Nigeria	800	23.8	34.0	43.2
Thailand	*771*	*22.3*	*30.0*	*47.7*
Tunisia	1140	17.2	33.5	49.3
Turkey	1120	20.5	33.9	45.6
Zimbabwe	629	13.6	34.3	52.1
Un-weighed Average (exc. Thailand)	869	21.0	30.5	48.5

The un-weighed average GDP per head of the 11 comparator countries was at US$ 869m, higher than Thailand, to reflect the strong growth prospects for the Thai economy over the next 10 years.

TOTs Consultants, D&N, in their data services study suggested Greece as an appropriate comparator country. This view was rejected by BT because, at US$ 3680, GDP per capita was almost 5 times that of Thailand. Turkey was much closer to Thailand on this criterion, and the structure of the Turkish economy also provides a better fit. For this reason, Turkey rather than Greece was included in the comparator list.

6.7.2 Residential Connection Growth

Telephone penetration, measured in terms of connections per household, in Thailand compares with the selected countries as shown below.

RESIDENTIAL PENETRATION COMPARISON
Connections per 100 households

	%
Dominican Republic	21.3
Tunisia	10.4
Ecuador	10.4
Turkey	10/0
Bolivia	9.9
Egypt	8.6
Thailand	*6.0*
Zimbabwe	5.9
Guatemala	5.6
El Salvador	5.6
Honduras	5.1
Cameroon	1.5
Nigeria	0.9
Average (exc. Thailand)	7.9

Penetration in Thailand was comparable with that in other countries with a similar GDP per capita.

6.7.3 Tariffs

The Residential line tariff of 50 Baht per month was equivalent to 3.85 per quarter, and was thus much lower than the corresponding BT tariff. The TOT Report and Accounts do not report profitability on a service by service basis, but it was unlikely that this tariff covered cost.

6.7.4 International Comparisons Business Connections

Telephone penetration, measured in terms of connections per US$ 1000 output in Thailand compares with the selected countries as shown below:

Turkey	19.98
Ecuador	10.82
Tunisia	10.01
El Salvador	9.80
Thailand	*8.10*
Egypt	5.41
Honduras	4.17
Zimbabwe	3.47
Guatemala	3.02
Bolivia	2.58
Cameroon	1.54
Nigeria	1.44
Dominican Republic	0.17
Average (exc. Thailand)	6.03

Penetration in Thailand was higher that in other countries with a similar GDP per capita - a likely consequence of the low tariff charged.

6.7.5 Tariffs

The Business line tariff of 50 Baht per month was equivalent to 3.85 per quarter, and was thus much lower than the corresponding BT tariff. The TOT Report and Accounts do not report profitability on a service by service basis, but it was unlikely that this tariff covered cost.

6.8 Japan International Co-Operation Agency (JICA) Study

The JICA master plan study on telecommunications development in Thailand (Dec. 1989) forecasts demand by two methods. One was based upon potential demand, called potential demand approach and the other based on expressed demand, called expressed demand approach.

Expressed Demand-----	----Hidden Demand----	-------Potential Demand---------
Existing Customers	Waiting Applicants	People who will subscribe to service as network service level is improved

The expressed demand consists of the existing customers and the waiting applicants. The potential demand consists of the expressed demand and those who were not included in the expressed demand but who were likely to have desires and financial capabilities of subscribing to the telephone service.

The potential demand was forecasted by estimating the number of potential residential customers and the number of potential business customers for each province.

6.8.1 Potential Demand Approach (Residential)

JICA developed a socio-economic model which was conceptually similar to CCITTs GAS 9 model. It was formulated on the basis of a household monthly income distribution to forecast the number of potential residential customers. The number of potential residential customers was obtained in three steps:

i) Estimation of household monthly income distribution

ii) Prediction of the number of households

iii) Calculation of the number of potential residential customers

6.8.3 Expressed Demand Approach

Expressed demand approach utilised two models on the basis of the data of existing customers and waiting applicants. The first model forecasted the number of customers in the Bangkok Metropolitan Telecommunications Area. The second model forecasted the number of customers in the provincial telecommunications areas.

6.8.4 The Results of Demand Forecast

In CCITT terminology, the Business Connections and Call models are both 'Econometric', and the Residential connections model is both 'Econometric' and 'Logistic', see figures 84 to 86. The methods they have used appear flawed as:

- They have not separately modelled Business and Residential connections

- They report serious multi-collinearity, i.e. correlation between variables, problems so have restricted themselves to simple models. The problem could have been overcome by fitting difference models (as used by BT). A price variable has therefore not been included in the model.

CCITT report that a 1% increase in GDP increases the customer base by 1.4%. The elasticity obtained from the BT models implies a smaller increase of 0.13%. The econometric technique that CCITT has used tends to give higher elasticities than the time series models used by BT and these are often interpreted as long-run effects. A conclusion that can be drawn is that the output elasticities in the BT

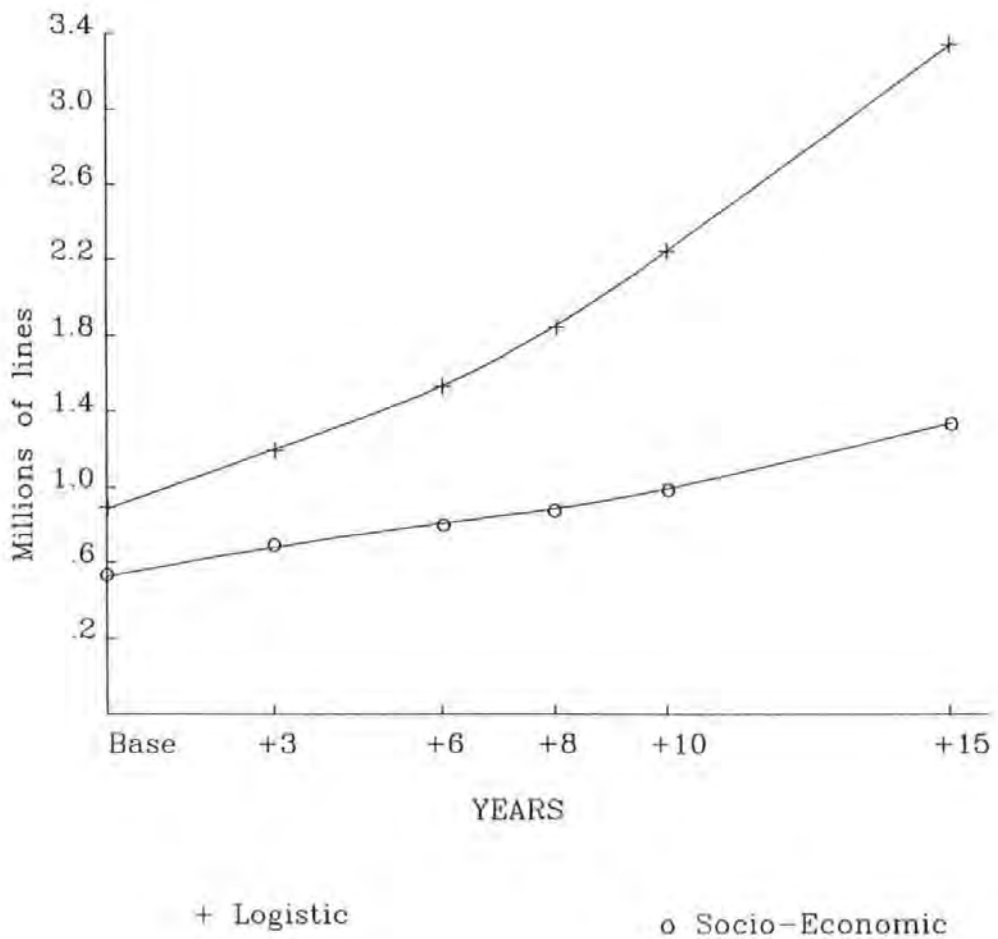


Figure 84 Demand Forecast in Metropolitan Bangkok using Logistic Model and Socio-Economic Model

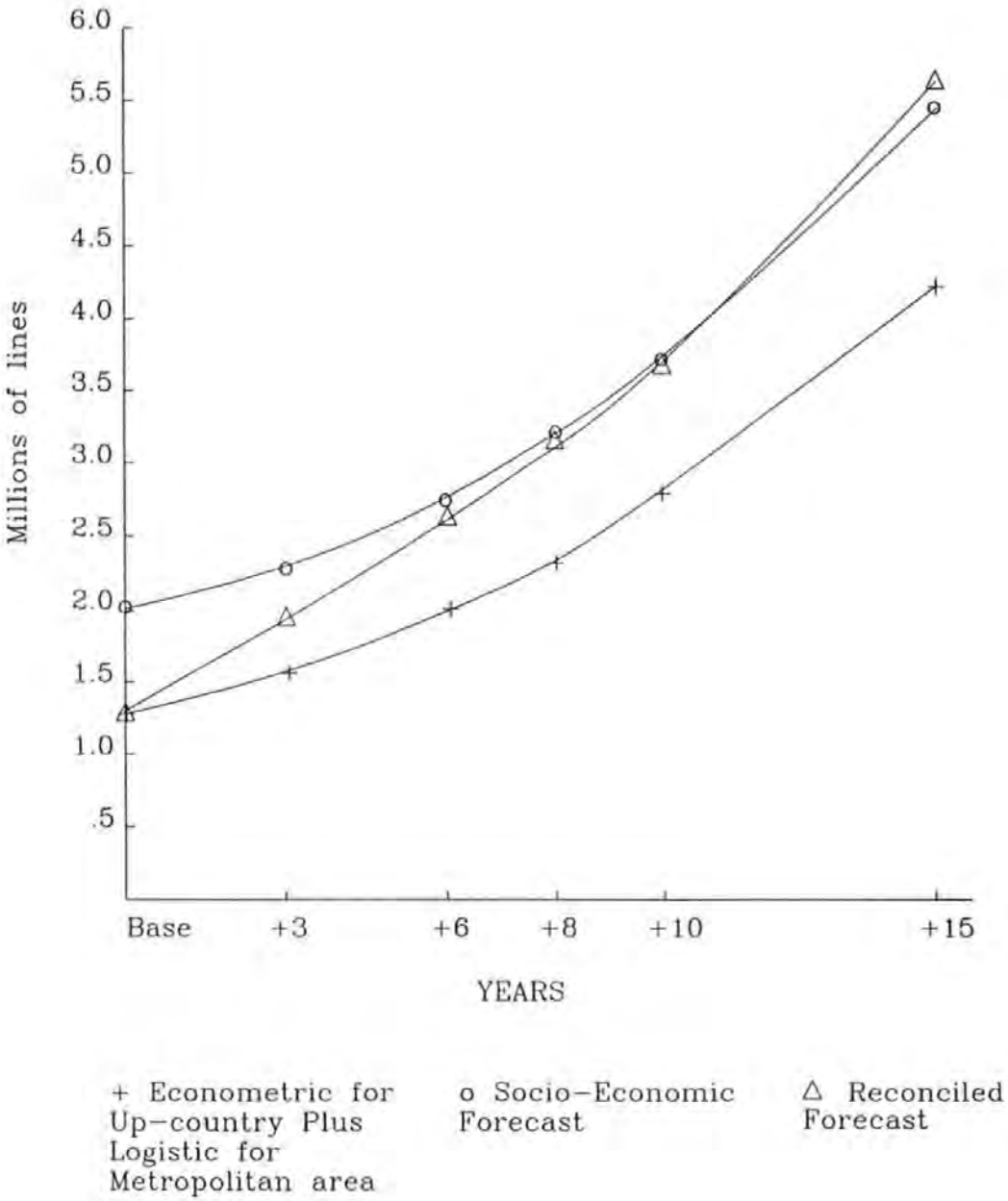


Figure 85 Demand Forecast for all Thailand

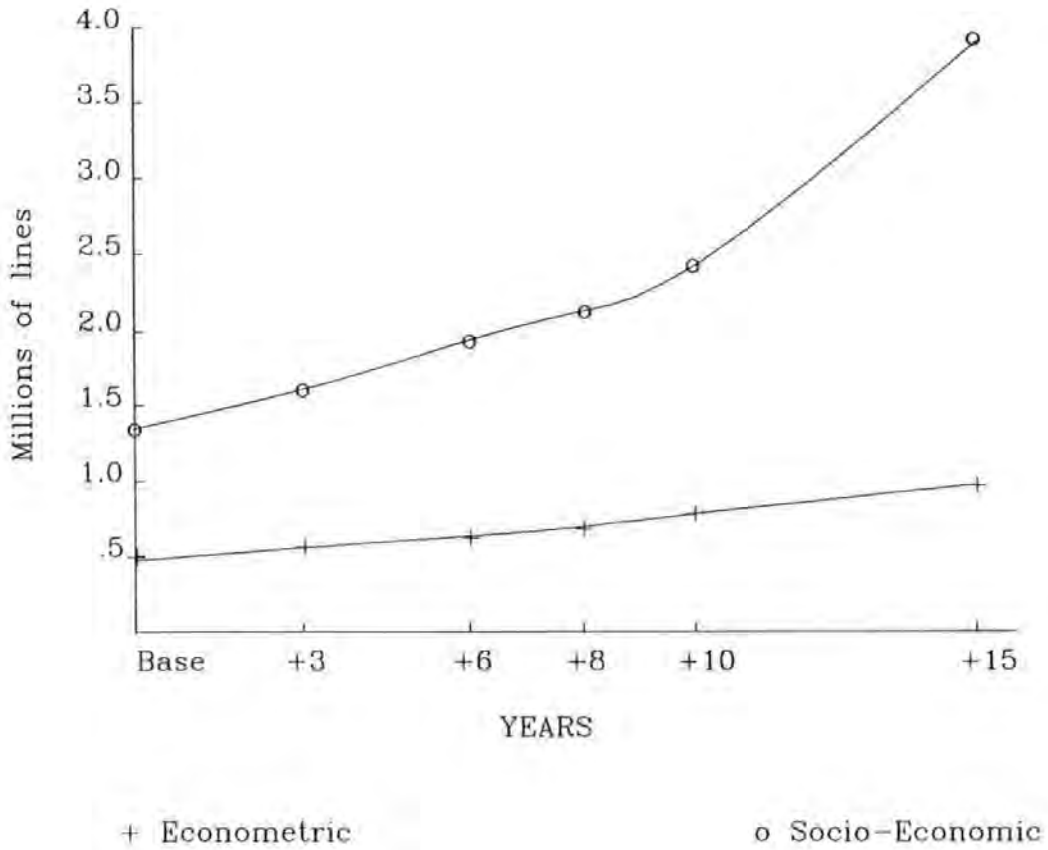


Figure 86 Demand Forecast for Up-Country Thailand using the Econometric Model and Socio-Economic Model

connections models may be low. The models can still be used for forecasting, however, as the trend elements in the model are correspondingly higher.

The CCITT forecasts start from an unidentified base year in which there were 1.3M customer lines. This would appear to be reported line capacity for 1987, although there were only 0.9M lines connected. BT therefore deducted 0.4M from the CCITT projections to reflect this unused capacity.

On this basis, the CCITT and BT central projections compare as shown in below.

YEAR	EXCHANGE LINE FORECAST			
	BT		CCITT	
	Provincial	Bangkok	Total	Total
1987	0.3	0.6	0.9	0.9
1990	0.4	0.9	1.3	1.3
1993	0.4	1.1	1.5	1.9
1995	0.5	1.3	1.8	2.3
1997	0.6	1.5	2.1	2.8
2002	1.0	2.3	3.3	4.1

This shows that short term projections from the two modelling approaches are similar, but that the BT models are more optimistic about medium term demand prospects. By 2002, the BT 'pessimistic' projection would show a total of 3.0M connections, so the CCITT forecast lies reasonably close.

The socio-economic method provides a larger magnitude forecast of demand in the provincial telecommunication areas, unconstrained historical trends which may have been influenced by TOT administration policy and tariff structures. On the other hand, the logistic method provides a larger magnitude forecast for the Bangkok telecommunication area, since it makes a better assessment of the demand for business services.

Since the use of different forecasting methods have resulted in different estimates of the potential demand there is a need to reconcile these results. A key consideration is the potential demand forecasted for the target year, using method socio-economic for the provincial telecommunication areas, and logistic for the metropolitan telecommunication area. This demand needs to be adjusted to account for a possible delay in providing the service.

JICAs forecasted potential demand was larger than the forecasted expressed demand in the provincial telecommunication areas. They stated that a reason may be that much potential demand has been discouraged registering as waiting applicants. Therefore, JICA considered the forecasted potential demand more appropriate because discouraged demand would very likely show up once network accessibility was improved.

JICAs forecasted potential demand was smaller than the forecasted expressed demand in the Bangkok Metropolitan Telecommunication Area. They stated that the potential demand by business customers in the BMA should properly be higher than the forecasted figure by the model and recommended further study.

Hence their report employs the forecast results of the potential demand approach for the provincial telecommunication areas and the forecast results of the expressed demand approach for the BMA.

7 Node Locations and Link Network

The BT network design for Bangkok chose four node locations each with three remote peripheral units. The node locations were:

Bang Yai

Kutapani

Don Maung

Pamprat Satran Phani

The nodes were interconnected by a star network based upon Don Maung.

TOT's design did not take into account a 'green field' situation and distributed their forecast of demand over existing nodes.

The Japanese study did not detail the location of the nodes or the network topology.

APPENDIX 3

REFEREED PUBLISHED PAPERS

1. Reynolds, P.L., Sanders, P.W. and Stockel, C.T.: "Uncertainty in Telecommunication Network Planning", Proceedings Summer Computer Simulation Conference, Boston, p.222-227, July, 1993

2. Reynolds, P.L.: "An Expert System Approach to Private Telecommunication Network Planning", BT Technol. Jour., Vol.11, No.4, p.41-50, Oct., 1993

3. Reynolds, P.L., Sanders, P.W. and Stockel, C.T.: "Uncertainty in Telecommunication Network Design", (Submitted for Publication) Expert System (UK), 1994

PROCEEDINGS OF THE 1993 SUMMER COMPUTER SIMULATION CONFERENCE

JULY 19-21, 1993
LAFAYETTE HOTEL
BOSTON, MASSACHUSETTS

TWENTY-FIFTH ANNUAL SUMMER COMPUTER SIMULATION CONFERENCE

Edited by
Joel Schoen
The MITRE Corporation

Conference Sponsor
The Society for Computer Simulation
SCS

UNCERTAINTY IN TELECOMMUNICATION NETWORK PLANNING

P. L. Reynolds
BT plc, London, England
P. W. Sanders and C. T. Stockel
University of Plymouth, Devon, England

Abstract

The automation of the Telecommunication Network Planning process to produce practical network designs poses a unique problem that needs to be overcome if it is to be implemented by Expert Systems. The problem is one of **Uncertainty.**

A network design has to be viewed over a relatively long period of time from the beginning of the planning period. It needs to make allowances for network growth, changes in technology and deviations from the original design data, eg demand, prices, availability of resources and hardware, etc. The uncertainty of this data must be taken into consideration with particular attention being paid to the identification of the most sensitive inputs, so that either different or more resilient designs can be drafted.

This paper reviews the network design process and indicates areas where 'uncertainty' has significant effects upon the eventual design. Current methods of dealing with uncertainty in Expert Systems are shown to be left wanting; their singularity in dimension coupled with a loss of their compilation means that the intrinsic richness of uncertainty values to add to the design process has been lost.

The graphical concept of 'Uncertainty Windows' has already been proposed by the authors (Reynolds et al. 1993). The technique works by virtue of its ability to retain the composition of the uncertainty and allows for better solutions to be obtained. In the paper cited above, the problem of synchronisation of a network was detailed and shows, in principle, how the Uncertainty Windows technique is employed. This paper reviews the technique and gives examples of the Windows in more general operation, in particular, how the various output windows are combined to produce the final practical solution.

The application of Uncertainty Windows to security systems is also considered to indicate the wider usage of the technique.

Introduction

Expert systems are now widely used in solving telecommunication problems. They can be generically placed into two groupings. The first are 'assist' models that do not produce optimised solutions themselves; instead they simulate the consequences of alternatives under the control of a network planner. The second are 'search' models in which a systematic search is performed, using algorithms, to find optimal solutions.

However, with real network planning problems it is not always possible to match the required theoretical conditions specified by network design algorithms without making many assumptions and simplifications that can invalidate the optimal result.

Expert Systems are unlikely to provide a universal solution to the network planning problem by themselves. A large portion of the problem still requires a quantitative and analytical approach. This is especially true after the overall problem has been divided into smaller pieces, which are small enough to be formulated into mathematical models. Thence a combination of heuristic and traditional programs should provide the best and most efficient solutions under the overall management of the expert system.

The Expert System breaks down a large network problem into its component parts, use different methods to solve each of them, and then combine these results to create an overall solution. An appropriate degree of coupling between sub-problems has to be maintained however, if the solutions for the sub-problems are to stay near-optimal for the global problem. In this context, the process of sub-dividing can be viewed as a heuristic approach to the overall network optimisation problem.

The network planning problem can be sub-divided into a number of well-defined problem domains:

Traffic Analysis;
Routeing;
Numbering;
Switching;
Transmission;
Signalling;
Synchronisation;
and Capital Analysis.

With an Expert System approach it is therefore necessary to drop the goal of finding the optimum network, and instead, settle for near-optimum designs obtained by iteratively solving the series of sub-divided problems. This

should not be cause for concern as the cost minimum is fairly flat, slight deviations from the optimum point are normally of little importance.

Uncertainty in Telecommunications Network Design

There are two fundamental areas where uncertainty impacts upon the integrity of a telecommunications network design. The first is that although the ideal solution changes with time, the initial design has to be based around a single time, normally at the beginning of the planning period. It is essential to allow for the uncertainty in the accuracy and values of predicted data; particularly identifying the most sensitive inputs, so that either different or more resilient designs can be drafted. A second area concerns the uncertainty that resides at the initial design stage. Here it is necessary to generate a 'source-to-destination' traffic matrix. The measurement, estimation or allocation of traffic values forms the base input to all network planning models. It is here where uncertainty, in both data and application, will have its greatest effect, as errors made at this stage are propagated throughout the network plan.

Modelling Uncertainty

There are currently four generic techniques for dealing with uncertainty in Expert Systems. Each takes one of the following forms:

Linguistic (Non-numerical);
Logical (Categorical and Fuzzy);
Statistical (Probabalistic and Baysian);
Ad-hoc (Confidence and Certainty factors).

Linguistic techniques use language to describe uncertainty, thus allowing a computer to emulate the expert. Utilisation of such techniques in an Expert System necessitates the conversion of the terms into numbers for their subsequent manipulation. Unfortunately, there is considerable variation in the way that different people interpret phrases and this interpretation is context dependant.

Logical techniques that utilise traditional logic have an inability to handle conflicting knowledge. This is due to its intrinsic monotonic nature. Fuzzy logic, with its truth values ranging between zero and unity overcomes this problem. However, in application it suffers from the same problems that exist with linguistic techniques.

Statistical techniques utilise a system of weights associated with data based upon the users' interpretation of the frequency with which the data would be true in a long series of trials. Baysian is a recursive technique in which the posterior probability becomes the new prior probability of data when it is reused. Probabilities tend to be misused as it is often not possible in practise to determine the probability of an event, rather a personal probability is specified. Different people can assign, quite legitimately, different probabilities to the same event.

Ad-hoc techniques, such as MYCIN, have been developed to solve particular problems. Whilst they work well for that particular arrangement, many Expert Systems reuse the techniques even though the mathematical premise of their use under such circumstances can be questioned.

A common significant failing of current techniques is that they are singular in dimension. They do not take account of the different attributes of uncertainty that need to be propagated in different ways through the Expert System. Indeed, some attributes of uncertainty may be regarded as mutually inclusive whilst others are mutually exclusive. The composition of the resultant uncertainty contains ' intrinsic richness' on the problem being solved, for example, those rules or data that contribute the most to the 'doubt' of the final design can be found if the resultant uncertainty can be decomposed.

Review of Uncertainty Window Models

In the paper by the authors previously cited, it was suggested that data, facts, rules and system generated (inferred) data found in network design problems be considered as having one or more 'degrees-of-integrity'. Data, facts and rules were regarded as requiring degrees for:

confidence;
reliability;
relevance;

whereas inferred data requires:

orientation;
intensity;
profile;

making a total of six degrees of integrity.

These degrees of integrity are represented graphically by a series of windows, which show them on either x or y axes as appropriate. The essence of the process of associating the windows with their relevant rules, data and facts was that it gave the network planner an insight into the uncertainties that have helped to produce the overall system design. It indicates which sources of uncertainty and which assumptions, rules, facts or data are critical for further investigation to improve upon the confidence of the overall design.

The windowing technique works by virtue of its ability to retain the composition of the uncertainty and its associated values, assumptions, etc. and allows for better solutions to be attained.

Data Uncertainty Window

Data uncertainty, as per conventional expert systems, is entered as a personal probability or belief in the data being correct.

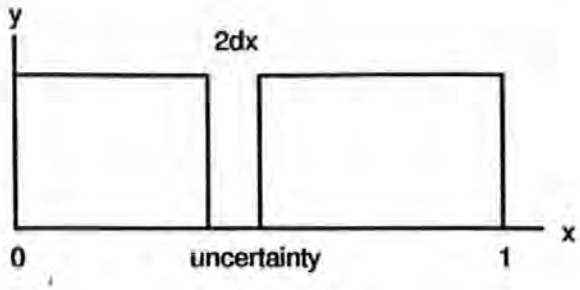


Figure 1

Fact Uncertainty Window

Fact uncertainty requires three degrees of integrity: confidence - similar to that of user data, shown on the x-axis; reliability - dependant upon their source, shown on the x-axis; relevance - dependant upon the application, shown on the y-axis.

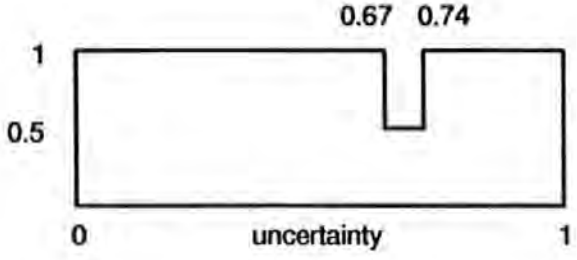


Figure 2

Rule Uncertainty Window

Rule uncertainty requires three degrees of integrity: heuristic content, shown on the x-axis; reliability - as with facts, shown on the x-axis; relevance - as with facts, shown on the y-axis.

Inferred Data Uncertainty Window

Inferred data requires three degrees of integrity: orientation - shown on the x-axis; intensity - shown on the y-axis; profile - shown on the x-axis.

Details of how these windows interact in a twenty rule system for synchronisation planning of a telecommunications network are given by Reynolds et al. (1993).

The Network Design Process

Teletraffic

Principles of teletraffic engineering state, inter alia, that the number of circuits required to serve a given type of traffic is not a linear function of the amount of traffic, in other words, if a given amount of traffic requires a specific number of circuits, doubling the traffic will not require double the number of circuits. Thus, it is preferable when optimising, to group together both voice and non-voice traffic to take advantage of this effect.

It is not always possible to 'Call Information Log' every site on the existing network in order to identify call levels and communities of interest. Indeed, in the case of 'green field' studies, there may be no site on which one can log. Under such circumstances various techniques are employed to estimate the expected traffic levels. These techniques include: penetration factors, where different types of building are assumed to have an approximate level of telephone penetration at an average calling rate; telephone bills, where call bills can be translated into units of traffic, called erlangs; job functions, where stereotypical functions have a traffic profile; user status within company or society, again with average traffic profiles; traffic per dollar of company turnover and numerous default values.

Each has an intrinsic uncertainty of varying degrees and an uncertainty according to the problem to be solved. Average and default values held as facts within the expert system will have associated questions regarding reliability, ie. quality of the source of the values.

Graph Theory

Typically network designs are based upon graph theory algorithms. However, the network design problem for reasonably sized networks and conditions is not solvable in polynomial time. In order to make the problem tractable, the only option available is to solve a related but less difficult problem. It takes the form of approximations on either the inputs or the outputs. In producing the minimum cost network design, the 'optimum' may not have been found, rather a very close to optimum or reasonable network being achieved. This may not increase cost by a large extent because of the normally 'flat' minimum of such optimisation. Having made compromises on the problem to be solved, it is a short step to compromising on the method to solve it.

The Expert System inteprets graph theory algorithms in an 'expert mode'. The analysis reduces the task of finding the optimum 'minimum cost' solution to one of near to optimum in a simpler practical manner. There are intrinsic uncertainties thus with each of the rules comprising the algorithm.

Transmission and Switching Network

A network designed by the use of graph theory, whilst giving the minimum cost for the topology, does not necessarily result in the optimum design for telecommunication networks. The practical trade-off between switching and transmission costs needs to be considered, as is the need for alternative routeing and maintenance of quality of service.

The uncertainty associated with this stage takes numerous forms. Firstly, the data itself will have intrinsic characteristics dependant upon its source. Secondly, assumptions have to be made concerning average costs per port and the division between fixed and variable port costs associated with switching. Thirdly, assumptions need to be made when calculating the cost of the access network, in particular the cost of the duct. The latter is either assumed to be negligible (unlikely) or zero if there is spare capacity in existing duct space.

Quality of Service

The grade of service of the network is automatically accounted for in the network design process. In summary, the grade of service is expressed as the percentage of calls lost or the proportion of time during which all the circuits on a route are engaged. It is possible to enhance the grade of service by taking account of the cost of failure. With private networks, for example, it is possible to provide a grade of service significantly different from that offered in the public network but at a correspondingly different price. Indeed, some private networks are provided on the basis that they are cheaper, albeit at a worse grade of service, than an equivalent service from the public network. It is not hard to see where uncertainty is encountered. The data collected in the initial market research will have to reflect the source of the information and hence its reliability. Scaling will need to account for the fact that what is one man's urgent is another's routine. Stereotypical characteristics can be employed but these will be averages and some users will not quite fit the pro-forma!

Call and Alternative Routeing Optimisation

The techniques for routeing optimisation have become more important in Great Britain as the process of deregulation offers the network planner a range of suppliers of circuits and network services.

Teletraffic principles dictate that when traffic between pairs of nodes is relatively small, a disproportionally high number of circuits are required for an acceptable grade of service.

A compromise is to provide a small number of direct links as a first choice with calls that cannot find a free direct link being routed via a transit switch. Provided that there is sufficient traffic to keep the direct links highly occupied, this will result in significant cost savings with little loss of efficiency. However, the true optimisation of alternative routeing with large networks is a very difficult algorithmic problem.

A near optimal routeing pattern can be obtained by heuristic methods. This entails routeing by an allocation process whereby a certain fraction of the call traffic from each switch is consolidated into a traffic group which is dependent upon its destination; each consolidated group being individually routed via its parent transit exchange.

Synchronisation

The overall performance of a private network is significantly affected by the synchronisation performance of the switches that comprise its nodes. Synchronisation difficulties in a network are often complex and manifest themselves as transmission impairments, such as error bursts and phase loss. Indeed, a significant portion of the transmission errors in a network can be attributed to synchronisation or more accurately to the lack of adequate synchronisation. The selection of the optimum centre of the network requires heuristic search methods.

The Plymouth Expert System

An Expert System shell has been developed that demonstrates the Uncertainty Window technique. It is written in PROLOG and dBase4; at present it runs on an IBM 386 model PC.

Communicating with the model and its results is critical to the success of the Expert System. One principal objective of the system has been to focus upon the maintenance of input as in conventional Expert Systems. User data is input as 'datum and confidence'. The window profiles of system-entered Facts and Rules are modified by their association with input data to produce concluding inferred data with its associated Uncertainty Window.

In principle, the process of generating the overall uncertainty profile of the 'solution' involves the superimposition of the window of the rule on the windows of its associated data, facts and inferred data. The resultant profile is generated by a Simpson's Rule numerical integration type of process. The effect can be visualised as the shadows produced after light has passed through all of the windows concerned in sequence.

Within the overall goal of finding the optimum network plan for a given set of inputs, it is necessary to solve a number of interrelated sub-problems. For example, prior to any dimensioning and costing of network components, it is necessary to define the communities of interest and traffic levels between these communities, ie. to develop a traffic matrix. At this stage of the planning process the technology of the network structure is irrelevant to the design process. It is necessary to separate the overall problem into a set of smaller problems but at the same time, maintain their relativity so that inputs and outputs can be made to flow in a natural process.

Describing the Problem

In a formal context, the overall problem can be described as contained within a problem frame represented as [PROBLEM FRAME]. Within the frame is a number of sub-problem domains represented as [PROBLEM DO-

MAIN] and a number of {SUB-PROBLEM DOMAINS} within which are the rules, facts, data and inferred data relevant to that sub-problem and can be represented as {FACTS, RULES, DATA, INFERRED DATA}. For each of these are their associated Uncertainty Windows represented as FACT(WINDOW).
Thus a simple problem would be described by:
[{Sub-Problem A (Fact A (Window), Rule A (Window), Datum A (Window)} {Sub-Problem B (Fact B (Window), Rule B (Window), Datum B (Window), Inferred Datum B (Window)}]
This representation is powerful as it is now possible to describe explicitly the interrelationship of the sub-problems. For example, consider the following hierarchical set of sub-problems:
[{sub-prob_a {sub-prob_b {sub-prob_d, sub-prob_e}}, sub-prob_c {sub-prob_f, sub-prob_g}}].
Here the output of sub-problem a feeds to both sub-problems b and c; the output of sub-problem b feeds to both the inputs of sub-problems d and e; the outputs of sub-problem c to the inputs of sub-problems f and g.
The current implementation status of the Plymouth Expert System can be represented at the coarsest level by:
[Geograpical Coverage {Penetration {Exchange {Direct, Graph, Synchro}, Matrix{Direct, Graph, Synchro}}, Building {Exchange {Direct, Graph, Synchro}, Matrix {Direct, Graph, Synchro}}, User {Exchange {Direct, Graph, Synchro}, Matrix {Direct, Graph, Synchro}}}].

The System in Use

Input

Three sub-problem domains dominate the input to the arrangement. They are called Penetration, User and Building. The penetration domain divides the area under examination into 250 by 250 meter blocks. The telephone density is calculated based upon the primary, secondary and tertiary utilisation of the square multiplied by its relevant penetration factor. Traffic calling rates are also calculated. The estimation of utilisation percentages will have a significant uncertainty as rough estimates are calculated based upon visual interrelation of percentage utilisation. This uncertainty is compounded by the multiplication of penetration factors that are based upon averages calculated from a number of network designs. These penetration factor uncertainties are pre-programmed in the expert system and have two dimensions: heuristic content; reliability. Relevance is not programmed at this stage and is introduced if and when the rule is activated outside the sub-program domain. The resultant inferred-data output has associated with it its set of uncertainty windows.
All buildings over three stories high or those buildings of one or two stories that are known to operate as businesses are accounted for in the Building problem domain. Here rules that account for telephones per square meter and dollar of turnover are used to estimate the number of telephones and associated calling rates. Uncertainty in the input data and the associated uncertainty windows of the rules and output inferred-data is similar to the penetration problem.
The final User problem domain takes account of special identifiable users. Here it is possible to be more accurate on the needs of each user. However, even here such users are likely to over-estimate requirements as the element of cost is rarely taken into account at the survey stage.

Interim

It is now necessary to change the telephone demand in each 250 by 250 square into a single source of traffic and a list of destination sinks. Two problem domains operate in parallel, each taking input inferred-data from User, Penetration and Building. The Exchange sub-problem takes a total of 784 squares and groups them in blocks of 16 to give a source/sink matrix. This is likely to be a good estimate of a traffic collecting local area. The Rule uncertainties and internal Fact uncertainty windows now combine with those of the input inferred-data.
The second problem domain generates the full traffic matrix from communities of interest data and a set of heuristic rules based upon experience from solving other problems.

Output

A series of problem domains now come into play. The final now calculates the minimum spanning tree to interconnect the sources and sinks, with boundary conditions that ensure the minimum number of hops is kept to four. This is called 'Graph' in our system. Another domain looks at providing direct routes, 'Direct'. Each domain has the aim of reducing cost and maintaining quality and grade of service. 'Synchronisation' selects the Primary, Secondary and Tertiary Master centers.
The final design comes complete with its associated combined Uncertainty Window. It is possible to back-track first to the data, rules or facts that contributed to the inferred-data window used as input. Thus if the output design is shown to be too unreliable in terms of the level of resultant variability it is possible to focus on those parts where uncertainty is greatest.
Figures 3 through 5 show the input data frames and output uncertainty windows from a typical session.

Another use of Uncertainty Windows

Another area of application for the Uncertainty Window technique is being examined at present. It covers the supervision and real-time monitoring of user profiles in large scale open computer system security. This involves

Map reference Vertical [ ]
Horizontal [ ]

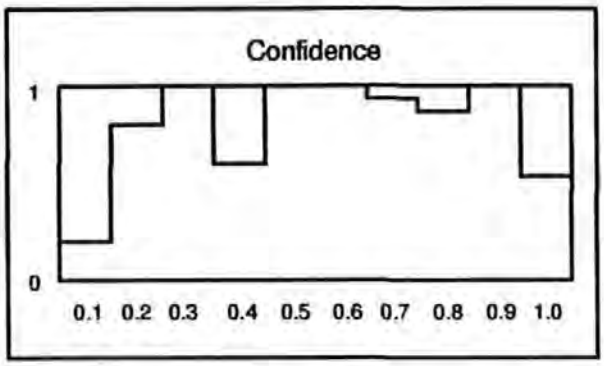
Primary Usage [ ] Percentage [ ]
Confidence [ ]

Secondary Usage [ ] Percentage [ ]
Confidence [ ]

Tertiary Usage [ ] Percentage [ ]
Confidence [ ]

Land Value [ ] Confidence [ ]
Population [ ] Confidence [ ]
Growth Rate [ ] Confidence [ ]

Typical Input Screen
Figure 3



Typical Output Window
Figure 4

The Primary master node Number is :-
1

The Area of the confidence window is :-
21

The confidence of the answer based upon probability theory is :-
0

The confidence factor CF associated with the answer is :-
9

Press any Key to continue

Output Text Screen
Figure 5

authentication of users to stop masquerade and limit unofficial access into large scale networked systems, which is becoming a serious problem. User identification numbers with passwords and perhaps an audit trail, provide a relatively simple and cheap form of protection but are considered weak in security terms. The inconvenience of remembering and frequently changing passwords should not be a necessary burden on the user. 'Smart' and 'dumb' plastic cards are more secure, but still require a PIN to be remembered and the card reader device can be very expensive. Ideally the computer terminal should automatically authenticate users by biometric means, eg. voice, image, fingerprint, signature, etc. with minimum effort from the user. However suitable devices for these methods are expensive and have wide tolerances of acceptance.

A different approach is to use **session supervision** based on services provided by a Security Management Center (SMC). Such a center may be provided by network operators for public key distribution and message certification activities involving non-repudiation security services, as defined in ISO 7498/2. Such a location or locations to which the user is 'parented' could also provide facilities for real-time monitoring of sessions, but in a completely secure manner for the session users.

The SMC could monitor, directly or indirectly, a number of attributes that make up a **user profile**. The keyboard activities and keyboard user characteristics can provide a significant input, with the normal sequencing of user events, the type and time of events, the location of the terminal, etc. giving additional information. The various factors with their variables can form the sub-domains of the windowing technique, to provide a final outcome of an authentication grading, which may change during the period of the session.

If the supervision control has 'doubts' , the attributes seem to produce too uncertain a result at the time, a more proactive stage can be made whereby simple questions (previously input to the SMC by the users individually and in secret during their initialisation of the system) can be sent to the user terminal under suspicion asking for the correct answer. Such questions as 'what is your Mother's maiden name?', 'when did you graduate?', etc. may be appropriate. The accuracy and delay of the response, and with reference to the system security, will indicate what action should be taken at the SMC.

Such an arrangement is being investigated for a hospital environment where authentication of users and data integrity are of paramount importance, but costs must be kept to a minimum.

Reference

Reynolds, P L, Sanders P W and Stockel, C T, 1993, Simulation, 60 (submitted for publication)

An expert system approach to private telecommunications network design

P L Reynolds

Network planners have used computer systems for many years now as an aid to the design of large telecommunications networks. Such systems are generally used to process the vast amounts of data that are presented as the input to the design process and carrying out the large number of calculations often required by the design algorithms. The paper reviews the current status of methods used in private and overlay network planning and shows that it is necessary to reduce a large network design problem into component parts, use different modules to solve them and then combine the results to create a total solution. Detail is given of the Plymouth Expert System, developed by the author, which is a realisation of a computersied process that uses heuristic rules of thumb, being based upon his experience of over 15 years of designing such networks. At each stage of the design process, focus is maintained upon the impact of uncertainty on the final design. In particular, it shows that current methods of dealing with uncertainty are not suitable in supporting the solution of these complex problems and a new graphical method, called 'uncertainty windows', is proposed. In addition, details of how the eventual design can be improved by the use of feedback under the direction of these 'windows' is given.

Throughout the paper the word 'heuristic' is used in regard to any method or technique obtained by expertise or 'rule of thumb' activities and used to improve the efficiency, or feasibility, of a problem-solving system. A heuristic algorithm may work well with a variety of problems and may fail on others.

1. Introduction

Initially, the paper gives a background to various computer models that can be used in the network design process, with an indication of their limitations. The private network design problem is then detailed as a series of sub-problems, each being discussed together with the heuristics necessary to solve them. The current methods of dealing with uncertainty within expert systems are then examined and a new graphical method more suited to telecommunications network design problems is reviewed.

Kai-li [1] categorises all methods of optimising procedures used in the network planning process into two distinct groups.

- The first group is classified as an 'expert assist' set of models. These models do not produce optimised solutions themselves, but compute or simulate the consequences of alternatives under the control of a skilled network planner. The planner, on the basis of experience or intuition, proposes a series of different plans or designs, inputs them into the computer, then evaluates and compares the results. A good overview of assist models is found in Roth et al [2] and Mantelman [3], whilst Ferguson et al [4] gives the details of an assist model used by Bell Canada.
- The second group is classified as a 'logical search' set of models. These models initially create a 'search space' that contains possible solutions to the problem; the computer then performs a systematic logical search by using theorems and heuristic algorithms (usually from graph theory) to find the optimal solution. Because this group of models is based on well-established algorithms, they should give the optimal (or close to optimal) result (see Grout et al [5] for an overview).

In practice, however, when confronted with real network planning problems, it is not always possible to match the required theoretical conditions required by the various algorithms. For those that can, often many assumptions and simplifications need to be made, causing the model to be quite different from the problem, thus invalidating the optimal result.

Tange [6], argues that it is not possible simultaneously to optimise all variables in the network planning process. Iterative methods, providing a number of sub-optima, are normally used giving an acceptable approach to an optimal network plan. Gupta [7] has suggested that network planning is more of an art than a science! He

argues that it is often impossible to describe factors such as accommodation, regulation and organisational issues as a mathematical model — but they can be simulated by an expert system.

Furthermore, the optimisation of large telecommunications networks is included in a family of problems that are regarded as being intractable, i.e. non-polynomial (NP) in complexity theory terminology. Such problems, though solvable in principle, are impossible to solve in practice when a network comprises a large number of switching nodes, link arrangements and routeing conditions. Penrose [8] states: '...problems in NP are regarded as intractable for a reasonably large n, no matter what increases in operational computer speed, of any foreseeable kind, are envisaged. The actual time that would be taken (to solve the problem) rapidly becomes longer than the age of the universe'. Thus the network planning problem, with its large number of variables often has no efficient search solution. It is usually necessary to make assumptions about either the input or output of the large problems and as such the use of heuristic methods of search and problem solution becomes essential.

2. Expert system approach

Expert systems are unlikely to provide a universal solution to the large network planning problem by themselves. A portion of the problem still requires a quantitative and analytical approach. Solutions become possible if the overall problem is sub-divided, with certain sub-problems solved by the use of mathematical models and others by heuristic means — so a combination of heuristic and traditional can provide the best and most efficient solutions under the overall management of an expert system.

Thus, it is necessary to divide and sub-divide the overall problem into smaller problems until some parts become mathematically tractable. An appropriate degree of coupling between the sub-problems has to be maintained, however, if the solutions for the sub-problems are to stay near-optimal for the global problem. In this context, the process of sub-dividing can be viewed as a heuristic approach to the overall network optimisation problem. It may be, therefore, necessary to drop the goal of finding 'the optimum network design', and, instead, settle for 'near-optimum designs' obtained by iteratively solving the series of sub-divided problems. This should not be cause for concern since the cost minimum is fairly flat and thus slight deviations from the optimum conditions are normally of little importance. The mechanism of sub-division is fundamental to the success of the process and will be problem-dependent.

3. Sub-problems to the private network planning problem

It is possible to sub-divide the private network planning problem as shown in Fig 1.

3.1 Teletraffic

Principles of teletraffic engineering [9] state *inter alia* that the number of circuits required to serve a given type of traffic is not a linear function of the amount of traffic. In other words, if a given amount of traffic requires a specific number of circuits, doubling the traffic will not require double the number of circuits for the same grade of service. The grade of service being expressed as the percentage of calls lost or the proportion of time during which all the circuits on a route are engaged. Thus, it is preferable when optimising a design to group together as much traffic as possible to take advantage of this effect.

A source/destination traffic matrix usually forms the base input to all network planning models. Often, it is

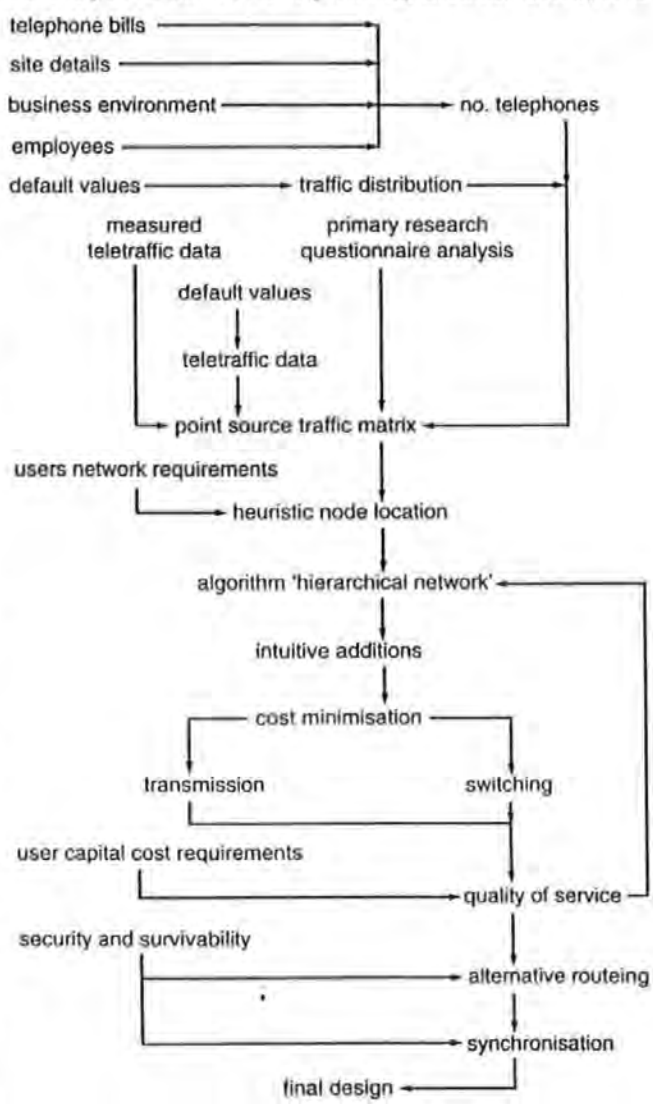


Fig 1 Heuristic network planning process.

not possible to monitor the call information at every site on the existing network in order to identify call levels and communities of interest. Indeed, in the case of 'green field' studies, there may be no site that can be monitored. Under such circumstances various heuristic techniques are employed to estimate the expected traffic levels in order to generate a 'point source traffic density' matrix.

Heuristic techniques for determining the traffic for the 'point source traffic density' matrix include:

- penetration factors — where different types of building are assumed to have an approximate level of telephone penetration at an average calling rate,
- telephone bills — where call bills can be translated into units of traffic (erlangs or call connect seconds equivalent),
- job function — where stereotypical functions have a traffic profile,
- user status within a company or society — again with average traffic profiles,
- traffic per £ of company turnover,
- numerous default values.

Each technique has intrinsic uncertainties of varying degrees and an additional uncertainity according to the problem to be solved. Average and default values held as facts within the expert system will have associated questions regarding reliability, i.e. quality of the source of the values.

Errors made at this stage are promulgated throughout the network plan and it is here where uncertainty, in both data and application, will have its greatest effect.

3.2 *Private branch exchange switch locations*

The 'point source traffic density matrix' now needs to be transformed into a structure that allows switch locations to be determined. Traditional methods, based upon graph theory, depend upon a source/destination matrix and are not 'usable' at this stage. A heuristic technique to find the optimum location of private branch exchanges (PBXs) has been developed being based upon a three stage process.

The first stage relies on the relationship between company turnover and telephone penetration in order to give a 'total' number of lines for each of the regions or sites that comprise the customer base. Secondly, from these totals it is possible to calculate the optimum number of switches needed to support this number of lines, based upon actual switch costs and capacities rather than traditional costs per port. Thirdly, these switches are distributed to a geographical location by a process of identifying descending 'centre of gravity' peaks of a three-dimensional plot of the 'point source traffic matrix' by progressively eroding the matrix peaks until the number of switches has been found.

3.3 *Interconnecting transmission network*

Typically private and overlay network transmission networks are designed using graph theory algorithms. The Esau-Williams [10] design algorithm can be radically simplified with the addition of boundary conditions, these not only produce results favourable to private network designs, with a resultant reduction of the calculation complexity, but also allow for its implementation in a heuristic way.

A typical private network design is equivalent to a two level 'hierarchical-net' (see Fig 2).

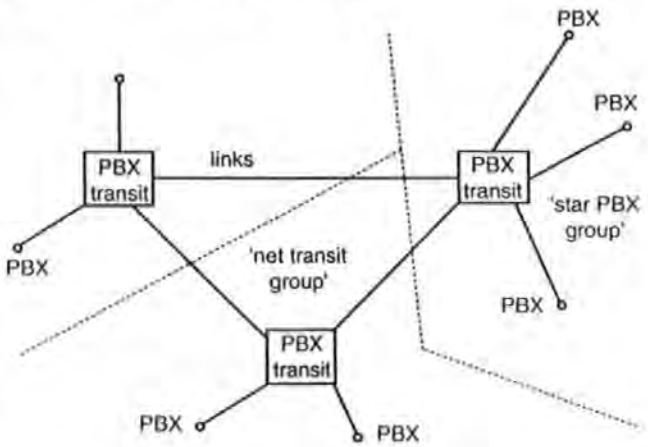


Fig 2 PBX hierarchical-net.

Two PBXs within the same transit group communicate via their common transit PBX. Traffic from a PBX to a peer in another group passes into the transit network, along a suitable route to the appropriate transit, then down through the service and supply network to the target PBX. In this way, the transits collect and distribute traffic around the network in an effective manner.

This structure is particularly effective since a large proportion of traffic generated within a PBX is likely to remain local to that exchange or one of its neighbouring exchanges. The multiple star arrangement at the lower level of the network deals with this well. The complete flexibility of the higher level, on the other hand, allows transits to be connected in the manner most appropriate to the characteristics and distribution of the traffic between them.

Thus it is necessary to construct a two-level hierarchical-star and then add additional inter-transit routes, which should result in a structure that can give the required network grade of service (GOS). The

following heuristic algorithm constructs a constrained hierarchical-net of a network with n PBX nodes and traffic weighted links.

The algorithm proceeds as follows:

(a) construct a minimum spanning tree of the form 'star' on each PBX node i=(1 to n) in turn,

(b) select that PBX acting as centre which gives the minimum cost 'star' tree,

(c) calculate cost T_{ij} for each pair (ij),

(d) do while $T_{ij} <> \infty$,

(e) find the pair (i^*,j^*) such that $T_{i^*j^*}$ is a maximum (the constraints are that PBX_{i^*} should not be more than two 'hops' from the centre and traffic capacity of links are not exceeded),

(f) if none of the constraints are violated by connecting i^* to j^* and disconnecting i^* from the centre, then do so; set $T_{i^*j^*} = \infty$ and go to (e) — otherwise set $T_{i^*j^*} = \infty$ and go to (e) (this gives a hierarchical-star configuration for PBXs and their transit).

A hierarchical-net is then constructed by the addition of links between level-two star networks so that the overall GOS is met:

(g) calculate GOS_{xy} for each pair (x^*,y^*), such that $GOS_{x^*y^*}$ is a minimum,

(h) do while $GOS_{xy} <> 0$,

(j) if the constraints are not violated by connecting x^* to y^*, then do so, and set $GOS_{x^*y^*}=0$ and go to (g). Otherwise set $GOS_{x^*y^*}=0$ and go to (g).

Calculating the network GOS is relatively simple; since the network takes the form of either a hierarchical-star or hierarchical-net, the worst case losses are additive.

3.4 *The optimum transmission and switching network*

A network designed so far, whilst giving the minimum cost for specific conditions, does not necessarily result in the optimum design for practical private telecommunications networks. The trade-off between PBX switching and transmission cost still needs to be considered, as does the need for alternative routeing and maintenance of the quality of service.

If the aggregate PBX switching cost of the network is greater than the aggregate transmission cost, then it is very likely that the elimination of one of the switches will result in a plan with lower real cost. This is because each PBX has a high fixed cost irrespective of the number of ports connected plus an incremental cost per port connected, whereas transmission is available in low-cost 30-circuit blocks, and interconnecting PBXs require non-customer facing capacity. The interconnection fibre optic circuits come with vast intrinsic capacity and are not pivotal in cost minimisation.

The decision as to which PBX to eliminate becomes one of costing the access network of each PBX switch and selecting the one that has the lowest aggregate access cost. This process continues until a point is reached where the aggregate transmission cost exceeds that of the aggregate PBX switching cost.

In each costing exercise, it is essential that the most economical combination of circuits be provided at each switch. This necessitates the bulking of the various traffic types on to a single route making full use of suppliers' circuit tariff alternatives.

The uncertainty associated with this stage usually takes numerous forms. Firstly, the data itself will have intrinsic uncertainty characteristics dependent upon its source. Secondly, assumptions will have to be made concerning the average costs per port and the division between fixed and port variable cost associated with switching. Thirdly, assumptions will need to be made when calculating the cost of the access network.

3.5 *Quality of service*

The grade of service of a network has already been accounted for in the network design process. It is possible to enhance the grade of service equation by taking account of the cost of failure. With private networks, it is possible to provide a grade of service significantly different from that offered in the public network, but at a correspondingly different price. Indeed, some private networks are provided on the basis that they are cheaper, although at a worse grade of service, than an equivalent service from the public network.

Quality of service allows for the fine tuning of the grade of service calculations such that a cost is associated to lost or repeated calls which is weighted against the cost of avoiding them.

The principle of quality of service is discussed by Littlechild [11], his analysis being directed at public network design. In summary, he defines the average congestion cost of a call, Q_c, as:

$$Q_c = w/\mu - L$$

where the user's time is valued at an average of £w per hour — this includes both the time cost and the nuisance value of waiting for a successful call, μ is defined as the capacity of the system under investigation and L is the mean arrival rate of calls.

It is possible to utilise this formula for private networks in a heuristic way. As part of the compilation

of the traffic plan, data is collected on the job function of the users and their status or level within each company. From this data, it is possible to define the quality of service required and hence the capacity of the associated access links to be provided.

It is not hard to envisage where uncertainty would be encountered. The data collected in the initial market research will have to reflect the source of the information and hence its reliability. Scaling will need to account for the fact that one man's 'urgent' is another man's 'routine'. Stereo-typical characteristics can be employed but these will be averages and some users will not quite fit the pro-forma.

3.6 Call and alternative routeing optimisation

The techniques for routeing optimisation have become more important in the UK as the process of deregulation offers the network planner a range of suppliers of circuits and network services.

As previously indicated, teletraffic principles dictate that when traffic between pairs of nodes is relatively small, it requires a disproportionately high number of circuits. A compromise is to provide a small number of direct links as a first choice, with the calls that cannot find a free direct link being routed via a transit switch. Provided that there is sufficient traffic to keep the direct links highly occupied, this scheme results in significant cost savings with little loss of efficiency. However, the optimisation of alternative routeing networks is so complex that it does not lend itself to algorithmic methods [12].

The earlier modified Esau-Williams algorithm can be enhanced by adding 'routeing by allocation' whereby the allocation of a certain fraction of the call traffic from each switch is consolidated into a traffic group dependent upon the destination, each consolidated group being individually routed to its transit exchange. Berry [13] has developed formulae for the design of high usage direct links between exchanges under these circumstances. It has been possible to simplify these to a set of five heuristic rules that give results within a small percentage of the accuracy of optimum. This tolerance on accuracy needs to be reflected in the overall uncertainty of the resultant design.

3.7 Synchronisation

The overall performance of a private network is significantly affected by the synchronisation performance of the PBXs that comprise the nodes of the network.

Synchronisation difficulties in private networks are often complex and manifest themselves as transmission impairments, such as error bursts and phase loss. Indeed, a significant portion of the transmission errors in private networks can be attributed to synchronisation, or more accurately to the lack of adequate synchronisation. Abate et al [14] have developed an expert system with '... over a thousand stringent rules' for synchronisation planning. No details are given in their paper as to how the system operates, nor of its rule base nor of its success in providing private network synchronisation plans. What is clear from the paper is that it is an 'expert assist' package only.

Reynolds et al [15] details a 'logical search' approach, using 'uncertainty windows', to help solve this synchronisation problem and forms one part of the expert system.

4. Network planning using the Plymouth Expert System

In conjunction with the Network Development Group, University of Plymouth, an expert system has been developed which supports the above heuristic approach to private network planning. In developing the expert system it soon became apparent that the current methods of dealing with uncertainty were unsuitable for network design.

4.1 Modelling uncertainty

There are currently four generic techniques for dealing with uncertainty in expert systems and, as shown in Fig 3, each takes one of the following forms:

- linguistic (non-numerical),
- logical (categorical and fuzzy),
- statistical (probabilistic and Bayesian),
- *ad hoc* (confidence and certainty factors).

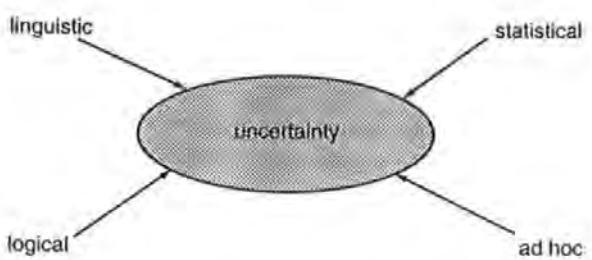


Fig 3 Dealing with uncertainty.

Linguistic techniques advocate the use of language to describe uncertainty thus allowing a calculus to mimic the expert. Utilisation of such techniques in an expert system necessitates the conversion of the terms into numbers for their subsequent manipulation. Unfortunately, there is considerable variation in the way different people interpret phrases and this interpretation is context-dependent.

Logical techniques that utilise traditional logic have an inability to handle conflicting knowledge, owing to

their intrinsic monotonic nature. Fuzzy logic, with its truth values ranging between zero and unity overcomes this problem. However, in application it suffers the same problems that exist as for linguistic techniques.

Statistical techniques utilise a system of weights associated with data being based upon the user's interpretation of the frequency with which the data would be true in a long series of trials. Bayesian is a recursive technique where the posterior probability becomes the new prior probability of data when it is reused. Probabilities tend to be misused, so it is often not possible in practice to obtain **the** probability of an event, rather, a **personal** probability is allocated. Different people can assign, quite legitimately, different probabilities to the same event.

Ad hoc techniques have been developed to solve a number of particular problems. Whilst they may work well for that specific problem many other expert systems reuse the techniques, even though the mathematical premise of their use under different circumstances can be questioned.

A common significant failing of current techniques when used in complex private network designs is that they are singular in dimension. They do not take account of the different attributes of uncertainty that need to be propagated through the expert system in different ways. Indeed, some attributes of uncertainty may be regarded as mutually inclusive whilst others are mutually exclusive. The composition of the resultant uncertainty contains intrinsic richness on the problem being solved; for example, those rules or data that contribute most to the 'doubt' of the final design can be found if the resultant uncertainty can be decomposed.

4.2 *Review of the 'uncertainty window' models*

The need for a new method in the construction of an expert system for telecommunications design has been detailed in a previous paper [16].

In that paper the principle of 'uncertainty windows' was proposed using a metaphor of light passing through a number of sub-problem modules the resultant 'shadow' giving a conclusion, with its associated uncertainty (see Fig 4). The data, facts, rules and system generated (inferred) data found in network design problems were considered as having one or more degrees-of-integrity. Data, facts and rules are regarded as requiring degrees for:

- confidence,
- reliability,
- relevance,

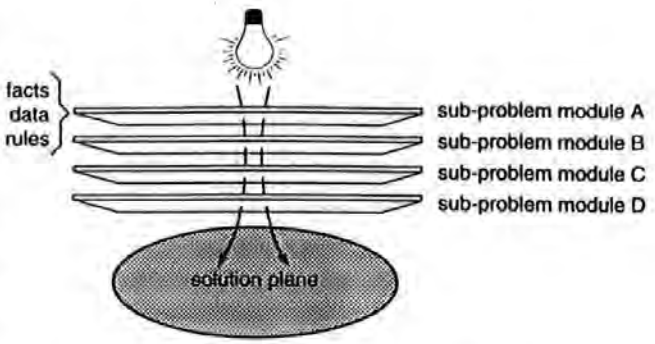


Fig 4 Uncertainty window metaphor.

inferred data requiring:

- orientation,
- intensity,
- profile,

a total of six degrees-of-integrity.

These degrees-of-integrity are represented graphically by a series of windows, which show them on either x or y axes as appropriate (see Fig 5). The essence of the process of associating the windows with their relevant data, facts and rules was that it gave the network planner an insight to the uncertainties that have helped produce the overall system design. It indicates which sources of uncertainty and which assumptions, data, facts or rules are critical for further investigation to improve upon the confidence of the overall design.

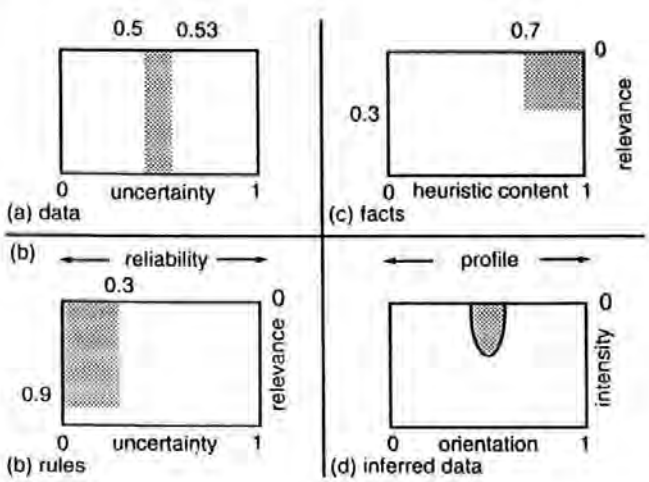


Fig 5 Uncertainty windows.

The windowing technique works by virtue of its ability to retain the composition of the uncertainty and its associated values, assumptions, etc, and allows for better solutions to be attained.

- Datum window

  Data is example specific, being entered by the user as in conventional expert systems — for input $x \pm dx$, where dx is the uncertainty value either side of the input value (Fig 5(a)).

  The model shows a window with a right orientation for less certain data and a left orientation for more certain data.

- Fact window

  Facts require three degrees-of-integrity — confidence, similar to that of user data, shown on the x-axis; reliability, dependent upon their source, shown on the x-axis; and relevance, dependent upon the application, shown on the y-axis (Fig 5(b)).

- Rule window

  Heuristic content of a rule is directly analogous to confidence with regard to facts and data. A rough and ready heuristic rule may often seem to give the correct answer, but without the rigour of mathematical proof can we be sure that it applies in many cases? The heuristic content must be reflected in the overall inferred data uncertainty profile, since the attributes of the rule are part of the expert system and invisible to the user. Rules of high heuristic content have right-oriented windows, and rules that are recognised laws have left-oriented windows.

  Reliability is a measure of the systems programmers' assessment of their confidence in the source of the rules, i.e. are they experts in the field? Reliability is directly analogous to its counterpart with facts, the x-axis or tolerance of the window being used.

  Relevance is directly analogous to its counterpart with facts, the y-axis or depth-of-cut into the window being used (Fig 5(c)).

- Inferred datum window

  The degrees-of-integrity for inferred data are — orientation, shown on the x-axis; intensity, shown on the y-axis; and profile, shown on the x-axis (see Fig 5(d)). The process of generating the uncertainty profile of inferred data involves numerical integration. In practice, the resultant profiles are complex; however, Fig 6 indicates how they can be interpreted.

  For details of the integration process the reader is referred to Reynolds et al [15].

4.3 *Plymouth Expert System*

As previously stated, within the overall goal of finding the optimum network plan for a given set of inputs, it is necessary to solve a number of interrelated sub-problems. For example, prior to any dimensioning and costing of network components, it is necessary to define the communities of interest and traffic levels between these communities, i.e. develop a traffic matrix. At this stage of the planning process the technology of the network infrastructure is irrelevant to the design process.

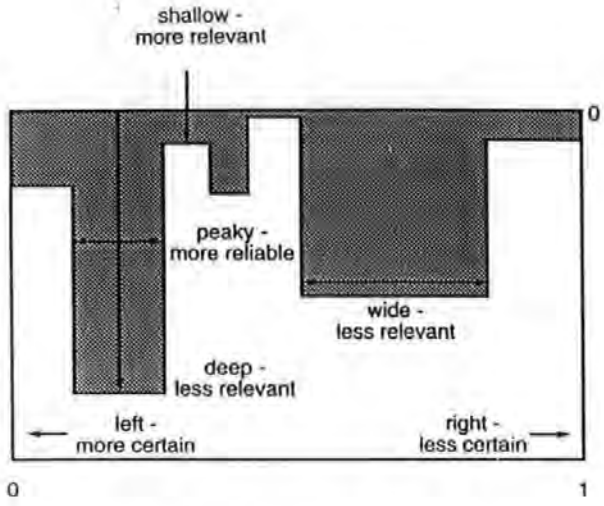


Fig 6 Inferred datum window.

Figures 7 and 8 show the modules that currently comprise the Plymouth Expert System and how they interrelate. Figure 8 also shows two forms of feedback being applied under the control of the 'uncertainty window'. The first form changes the format of the rule 'uncertainty window' when the expert system is used against real network problems; the second picks the most sensitive inputs.

Three problem domains dominate the input to the expert systems — penetration, users and buildings; Fig 9 shows two of these. The penetration domain divides the area under examination into 250×250 metre blocks. The telephone density is calculated based upon the primary, secondary and tertiary utilisation of the block multiplied by its relevant penetration factor. Traffic calling rates are also calculated.

The estimation of utilisation percentages will have imparted significant uncertainty as rough estimates are calculated based upon visual interrelation of percentage utilisation. This uncertainty is compounded by the multiplication of penetration factors that are based upon averages calculated from a number of network designs. These penetration factor uncertainties are preprogrammed in the expert system and have two dimensions — heuristic content and reliability. Relevance is not programmed at this stage and is introduced if and when the rule is 'fired'

- direct routes
- switches
- connectivity (transmission)
- transit network
- traffic matrix
- input data
  - -buildings
  - -users
  - -penetration
- synchronisation

Fig 7 Plymouth Expert System sub-problem modules.

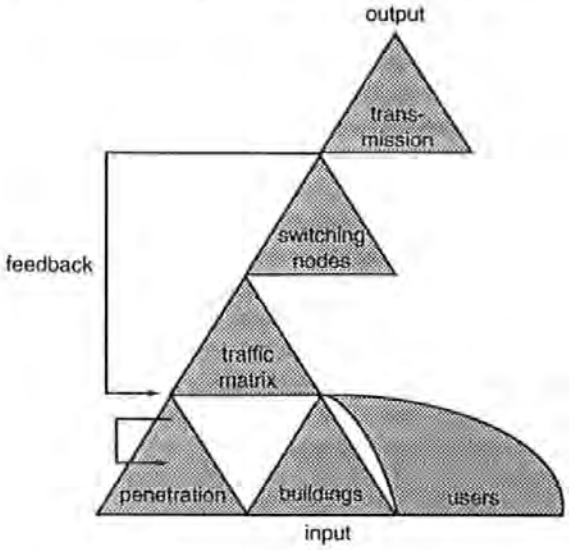


Fig 8 Outline of the Plymouth Expert System.

outside the sub-problem domain. The resultant inferred-data output has associated with it its set of 'uncertainity windows'.

All buildings over three stories of those buildings of one and two stories that are known to operate as businesses are accounted for in the building problem domain. Rules that account for telephones per square metre and pounds sterling of turnover are used to estimate the number of telephones and associated calling rates. 'Uncertainity windows' of the rules, input data and output inferred-data are treated as for the penetration problem.

The user's problem domain takes account of special identifiable users. However, even here such users are likely to over-estimate requirements as the element of cost is rarely taken account of at the survey stage.

It is now necessary to combine the telephone demand in each 250×250 metre block into a single traffic source and destination sink. Two problem domains now operate in parallel, each taking input inferred-data from user, penetration and building. The 'traffic' domain takes up to a total of 144 blocks and groups them as a local source/sink as this seems to give a good estimate of a 'local traffic collecting area', i.e. collection radius of 2 km. The rule and fact 'uncertainty windows' now combine with those of the input inferred-data. The second parallel problem domain, 'transit', generates communities of interest data from a set of heuristic rules based upon experience from other problems and calculates the total number of PBX switches required. The output from both these domains is used to generate a three-dimensional graph of the 'point traffic matrix' from which the location of the PBX switches is found.

A series of problem domains then calculates a basic topological design. Firstly, 'connectivity' constructs a minimum spanning tree to interconnect the source and sinks, with boundary conditions ensuring that the maximum number of hops is kept to four. The 'switching' domain looks for trade-offs between switches and transmission. Another looks at providing direct routes, called 'direct'. Each has the aim of reducing cost and

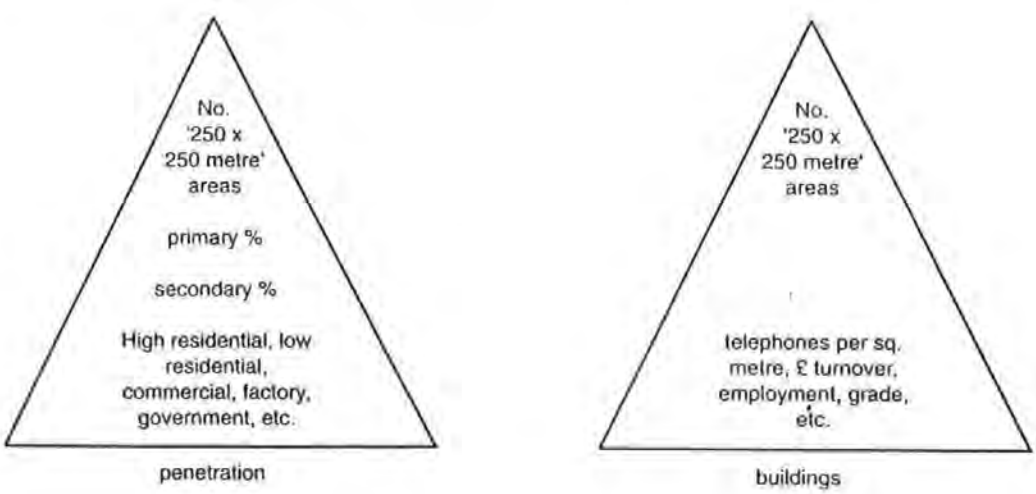


Fig 9 Sub-problem example.

maintaining quality and grade of service. Finally 'synchronisation' selects the primary, secondary and tertiary master centres.

The final design comes with its associated combined 'uncertainty window'. It is possible to backtrack first to the windows that made up this profile and ultimately to the data rule of fact that contributed to the inferred-data window used as input data. Thus is the output design is shown to be unreliable it is possible to focus on those parts where uncertainty is greatest.

In its present implementation the Plymouth Expert System groups together ranges of uncertainty. This facilitates speed of operation in the development of the system by speeding up the analysis process. It is a simple task to refine the graduation of the confidence ranges. For comparison, associated with the confidence windows are the results of both confidence factor and Bayesian analysis.

5. Conclusions

Conclusions can be drawn in two areas — the validity of utilising heuristic network planning methods and dealing with uncertainty.

Traditional network planning tools are removed from the practical world for which they were developed. They often ignore the significant uncertain and statistical nature of the input data; they use data taken from a fixed point in time to solve a time-variable problem and cost formulae tend to be on an average per line or port. A heuristic approach has been shown to produce acceptable results in a fraction of the time taken by traditional methods.

Current uncertainty calculus is not suited to the solution of complex problems such as network design. They are singular in dimension and do not facilitate feedback in order to improve the design. Facts, rules and data need to be treated differently. The uncertainty window is able to handle these different characteristics. The windows also present the user with a view of the whole uncertainty picture that allows for more effective perceptional enhancement of the solution in a computationally cheap and effective way, i.e. use of picture and the eye.

Acknowledgements

This paper outlines the research conducted by the author in association with Peter Sanders and Colin Stockel of the University of Plymouth. Peter Sanders is also investigating the application of uncertainty windows to real time security monitoring.

References

1 Kai-li K: 'Expert systems in telecommunications network planning and design', Proc 1st Int Con App AI in Eng, 2, pp 1161—1164 (1986).

2 Roth P F, Mohammad I and Moufiha H T: 'Simulation: a powerful tool for prototyping telecommunications networks', Simulation, 58, pp 78—82 (1992).

3 Mantelman L: 'AI carves inroads: network design, testing and management', Data Comms, pp 106—123 (July 1986).

4 Ferguson I A and Zlatin D R: 'Knowledge structures for communications networks design and sales', IEEE Nets (USA), 12, No 5, pp 52—58 (1988).

5 Grout V M, Sanders P W and Stockel C T: 'Practical approach to the optimisation of large switching networks', Comp Comms, 10, No 1, pp 30—38 (1987).

6 Tange I: 'General considerations by CCITT working party XIII/2 on the use of computers for network planning', Tele Jour ITU, 38, No 11, pp 737 (1971).

7 Gupta V P: 'What is network planning', IEEE, Comm Mag, 23, No 10, pp 10—16 (1985).

8 Penrose R: 'The emperor's new mind', Oxford University Press, pp 138—146 (1989).

9 Bear D: 'Principles of telecommunications traffic engineering', Peter Peregrinus Ltd (1976).

10 Esau L R and Williams K C: 'A method for approximating the optimal network', IBM Sys Jour, 5, No 3, pp 142—147 (1966).

11 Littlechild S C: 'Elements of telecommunications economics', Peter Peregrinus Ltd, Chapter 11 (1979).

12 Pratt C W: 'The concept of marginal overflow in alternative routeing', Aust Tele Res, 11, pp 76—82 (1967).

13 Berry L T M: 'An explicit formula for dimensioning links offered overflow traffic', Aust Telecom Review, 8, Part 1, pp 13—17 (1974).

14 Abate J E et al: 'AT&T's new approach to the synchronisation of telecommunications networks', IEEE Com Mag, 27, No 4, pp 35—45 (1989).

15 Reynolds P L, Sanders P W and Stockel C T: 'Uncertainty in telecommunications network design', Simulation, 60 (to be published) (1993).

16 Reynolds P L, Sanders P W and Stockel C T: 'Uncertainty in telecommunications network planning', Society for Computer Simulation Conference, Boston, pp 222—227 (July 1993).

Paul Reynolds joined BT in 1968 as an apprentice. He gained his Bachelor's degree from Plymouth Polytechnic in 1978 and moved into practical private telecommunications networks with BT's Tallis Consultancy, designing private networks for many multinational companies. Later, while in the overseas division, he was involved in designing mobile communications networks for Jordan, Thailand, Malta and Greece and telephone networks for Trinidad, India, Pakistan and Thailand. He has close associations with the Network Development Group (University of Plymouth) where a centre of excellence is being developed in Telecommunications Network Planning and Design. He is presently working in BT's procurement directorate.

Uncertainty in Telecommunication Network Design

P. L. Reynolds, P. W. Sanders
and C. T. Stockel

Network Research Group
Faculty of Technology
University of Plymouth
Drake Circus
Plymouth, PL4 8AA
Devon

Paper submitted for publication to

Expert Systems
The International Journal of Knowledge Engineering

published by Learned Information Inc.
143 Old Marlton Pike, Medford, New Jersey, NJ08055

1. Abstract

The design of telecommunication networks is a complex activity involving a vast amount of input data and numerous design rules. Much of that data is estimated and many of the rules are based upon experiental knowledge. The use of current uncertainty techniques in expert systems is not suited to the complexity of telecommunication network design. Having investigated the existing uncertainty problem, it is believed that a more rational method has been found. The paper presents the "Uncertainty Window" technique and its testing on an aspect of network design, that of network synchronisation.

The essence of the uncerainty process is one of associating the uncertainty windows with their relevant rules, data and facts, which results in providing the network designer with an insight into the uncertainties that have helped produce the overall system design. It indicates which sources of uncertainty and which assumptions are critical for further investigation in order to improve upon the confidence of the overall design. The windowing technique works by virtue of its ability to retain the composition of the uncertainty and its associated values, assumptions, etc. and allows for better solutions to be obtained.

2. Introduction

The prime objective of telecommunication network design is to provide the relevant equipment, at the appropriate place and at a reasonable cost, in order to satisfy the expected demand and give an acceptable grade-of-service. The complexity of the design process has resulted in the development of a number of optimization computer packages (Grout, 1988). Unfortunately, such packages often ignore the practical problems, ranging from constraints on the availability of equipment to the particular characteristics of the community to be served. This, together with a lack of accuracy in the data necessary for the design process, can remove such packages far from the pragmatic environment in which their authors wish them to be employed.

In circumstances where data are missing and practical restraints are imposed on the component parts of a network, "rules-of-thumb" are necessary, and in this case an expert system approach can prove to be more useful to the design engineer. Attempts have already been made to utilise expert system techniques in the design of telecommunication networks. An overview has been given by Lees (Lees, 1989); Salasoo (Salasoo, 1991) describes a Bellcore system for supporting the design process by integrating knowledge from several databases to guide the designer. Solo (Solo et al., 1990) has developed an investment tool which combines technical details about network structures with managers' subjective assessments, so that the network evolution can be examined against differing assumptions regarding demand and facilities. Kamimura (Kamimura et al., 1989) describes two NTT models which can be used for the comparison of public and private network services in terms of cost. However, none of these above systems take a holistic approach to the design process.

3. Uncertainty

A problem associated with network design is that although a solution changes with time the design has to be viewed from a single moment, normally at the beginning of the planning period. In practice, the design engineer looks forward over a period to the design date, which allows for network growth and compensation for design errors, eg demand, prices, availability of resources and hardware. One needs to allow for the uncertainty of this future data; particularly identifying the most sensitive inputs so that either different or more resilient designs can be drafted.

Uncertainty has been given a wide interpretation within expert systems and appears to be used whenever reasoning by strict logical implication is not considered possible. This may be due to inadequacies in the knowledge-base, insufficient or unreliable data on the subject of interest or because of stochastic relationships between propositions. Uncertainty may be about Quantity; eg the slope of a forecasting curve to the design date or may be about Structure; the shape of the forecasting curve to the design date.

The variety of types and sources of uncertainty along with a lack of detail differentiation of the form of uncertainty generates considerable confusion. The authors consider it important to distinguish clearly between the different types and sources of uncertainty, since their treatment and propagation through the expert system needs to be treated in different ways, introducing a need for a multi-dimensional method of dealing with uncertainty. Since these aims are practical rather than philosophical, the network design application allows one to avoid the many controversies that this topic can engender.

4. Modelling Uncertainty

A review of some of the many areas that use non-categorical knowledge including Market Research (Cartes-Rello and Golshai, 1990), Medical Diagnosis (Hughes, 1989) and Engineering (Shiraishi et al., 1989) plus more general surveys by Zwick (Zwick and Wallsten, 1989), Gaag (Gaag, 1990), Murphy (Murphy et al., 1990), Guan (Guan et al., 1990), Sheridon (Sheridon, 1991) and Spiegelhalter (Spiegelhalter, 1989) have shown that, in essence, the present state-of-the-art modelling of uncertainty takes one of thefollowing forms.

4.1 Non-numerical techniques advocate the use of linguistic terms to describe uncertainty which, it is argued, allows a calculus to mimic the expert better and provide a more natural explanation. Utilization of such techniques within Expert Systems necessitates the conversion of the linguistic terms into numbers for their subsequent manipulation. Unfortunately, there is considerable variation in the way different people interpret phrases and this interpretation is context dependent.

4.2 Categorical techniques utilise traditional logical techniques. They are monotonic in that the addition of new information cannot invalidate any previous deductions and hence expert systems utilizing a logical base have an inability to handle conflicting knowledge.

4.3 Probabilistic modelling utilises a system of weights associated with data by the users' interpretation of the frequency with which the data would be true in a long series of trials. That means it is a theoretical property of an infinite number of trials, rather than a single event.

4.4 Ad-hoc techniques have made use of a Confidence Factor, traditionally a real number in the range -1.0 to 1.0. It indicates the certainty with which each fact, datum or rule is believed. Positive or negative confidence factors indicating a predominance of confirming or opposing evidence respectively. The numerical techniques for propagating confidence factors and combining evidence are dependant upon the individual expert system design.

4.5 Bayesian inference is a recursive statistical technique where the posterior probability become the new prior probability of data when it is reused in a computation to give a new posterior value, given new evidence.

4.6 Fuzzy logic allows truth values to take numerical values between zero and unity, ie they can take the form true, not true, very true, more or less true, etc. and the rules of inference are consequently approximate, rather than exact. Fuzzy set modelling and manipulation utilizes the basic axioms of traditional set theory.

4.7 Shafer-Dempster modelling makes use of probability techniques utilizing "upper and lower" boundaries. These boundaries correspond to all possible eventualities concerning the rule base. Thus it should be possible to model a wide range of uncertainty in this way. Shafer-Dempster techniques are similar to Baysian but where the former focuses on the facts and data bearing upon a rule, the latter focuses upon the result of the rule operation.

5. Limitations of Existing Models

Current techniques for dealing with uncertainty are ad-hoc, crude and singular in dimension. These methods are sufficient for a rank ordering, which is the current practice, because order statistics are more robust than interval statistics. Techniques are needed that can weigh data along with costs, benefits and risks. It does not matter if a resultant network design has a 75 per-cent chance of being analog if there is also a 10 per-cent chance that there is a need to transport digital data; so a rank order of likelihood of hypotheses is not useful in this case. However, numerical methods pose problems because of:-

1) Lack of Quantitative Data

It may be that there is either insufficient past data or insufficiently confident experts to assess anything other than an approximate or imprecise figure.

2) Alien Representation to the User

It is more natural for naive users to think of uncertainty in terms of **probable, possible, likely,** etc. rather than numerical probability representation. Does the probability of say 0.67 of an answer being correct mean much to the non-mathematical user? Even if the statistician's viewpoint that **probability** is

probably (sic) the best way to represent uncertainty is correct, it is still not necessarily the best means for communication with most people.

3) Misrepresentation of Probabilities
Although most statisticians would support the use of probabilistic reasoning to model the chance effects of underlying hypotheses, many would reject the use of probability to express uncertainty about events that could not be regarded as either having been generated by some random process or from the development of some **information** as time progresses. They argue that there is no such thing as **the** probability of an event: different people, or the same person at different times, may legitimately assign different probabilities to the same event. That is, it is a **personal** probability rather than " the probability".

4) Differentiation between Ignorance and Uncertainty
Clark (Clark, 1990) shows that it is difficult for one to distinguish uncertainty from ignorance. For example, if a rule suggests that either conclusion 1 or conclusion 2 is best but does not provide any basis for choosing between them, a numerical approach will select the conclusion with the higher probability.

5) Insight to the conclusions
O'Leary (O'Leary and Kandelin, 1988) has said that for numerical systems it is necessary to express the strength of a belief as a number, as all numerical calculations compute aggregate numbers and keep no record of divergence in opinion (Henkind and Harrison, 1988).

6) Statement types
Uncertainty methods are used uniformly for the different kinds of statement – facts, rules and data. It is important to distinguish between the distinct statement types since they need to be treated in different ways.

7) Dimension
Uncertainty techniques are singular in dimension. They do not take account of the different attributes of uncertainty that need to be propagated through the system in different ways. Indeed some attributes of uncertainty may be regarded as mutually inclusive whilst others are mutually exclusive.

8) Inconsistency
Magill (Magill and Leech, 1991), Castillo (Castillo and Alvarez, 1990) and Garbolino (Garbolino, 1987) have shown that the implementation of numerical uncertainty techniques do not strictly follow the theories upon which they were based.

6. A way forward

It is proposed that the data, facts, rules and system generated (inferred) data found in network design problems be considered as having one or more **Degrees-of-Integrity.** Data, facts and rules are regarded as requiring degrees for:

1) confidence;
2) reliability;
3) relevance;
and inferred data requiring:
4) orientation;
5) intensity;
6) profile;
for a total of six degrees-of-integrity.

These degrees-of-integrity are represented graphically by a series of **windows**, which show them on either x or y axes as appropriate. This graphical representation technique allows the non-expert users of the system to have an insight into and appreciation of the uncertainties that have helped to produce the overall system conclusion. It indicates which sources of uncertainty and which assumptions or rules are critical for further investigation to improve upon the confidence of the answer. The insight will be qualitative in nature even if the expert systems from which they derive are quantitative. The only previous work known by the authors on the graphical communication of uncertainty is that of Ibrekk and Morgan (Ibrekk and Morgan, 1987). No work is currently known on the representation of its multi-dimensional attributes.

7. The Models.

The graphical concept of **windows** modelling data, rules, facts and inferred-data is now introduced and simple examples of their use shown. The window technique works by virtue of its ability to retain the composition of the uncertainty and allows for better solutions to be attained.

7.1 Data (Extrinsic to the expert system)

Data are example specific, being entered by the user as in conventional expert systems. For input x dx, where dx is the uncertainty value either side of the input value, we have:

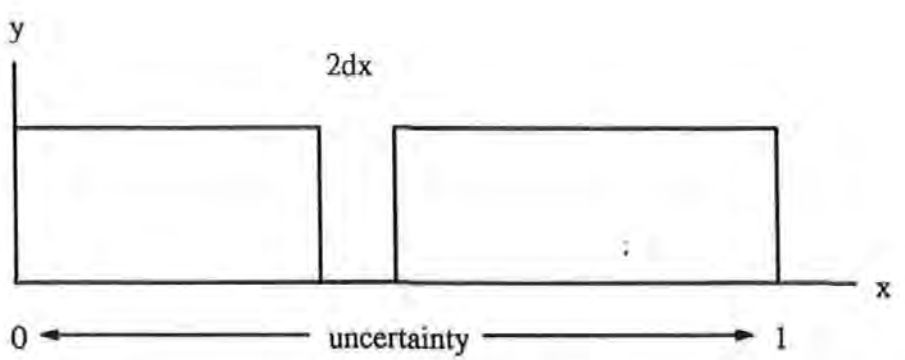


Figure 1

The model shows a **window** with a right orientation for less certain data and a left orientation for more certain data.

7.2 Facts (Intrinsic to the expert system)

Facts require three degrees-of-integrity:
confidence - similar to that of user data, shown on the x-axis;
reliability - dependent upon their source, shown on the x-axis;
relevance - dependent upon the application, shown on the y-axis.

reliability is derived from the quality of the source, ie is it from an expert, engineer or layman? There must exist a high correlation between reliability and confidence; indeed reliability can be regarded as "confidence in source". Thus the more confident that we are in the source, the less spread we have in the representation of confidence and this spread can be considered as a "tolerance in the confidence".

Taking the example in fig. 1, if our expert gave a confidence value of 0.3 (and hence an uncertainty value of 0.7) and we have a 90 per-cent confidence in the expert, then we have a spread in the confidence value of 5 per-cent (ie 0.035) as shown in fig. 2. This is analogous to a mean and standard deviation were the distribution to be Gaussian but this is not necessarily the situation, it could be any form of distribution.

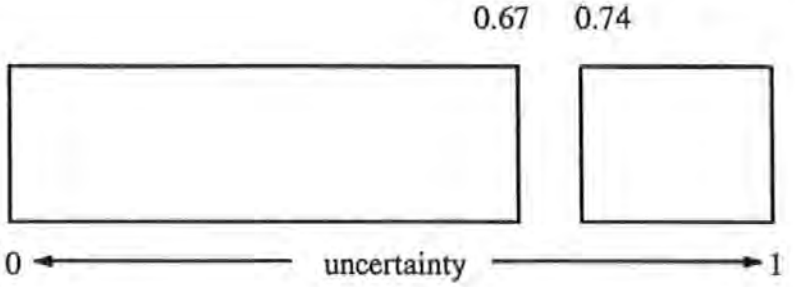


Figure 2

The same confidence given by a novice in the field could result in only a 10 per-cent confidence in the source, thus resulting in the window shown in fig. 3, of 0.70 0.32, with truncation of 1.02 to unity.

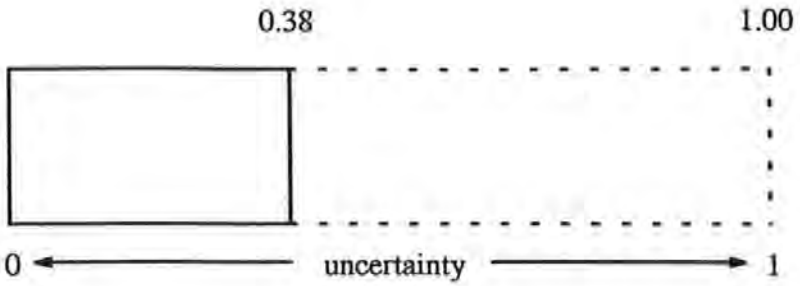


Figure 3

Figure 3 shows that the wider the window, the less confidence we have in the source of the fact and the more to the right the window, the less confidence we have in the value of the fact itself.

relevance is represented by the 'depth-of-cut' into the window. Totally irrelevant facts, no matter how certain the source or how confident we are in the source, have no part to play in the overall confidence of inferred data and will have a depth-cut of 100%. Highly relevant facts will have a shallow depth-cut and contribute more to the profile of the resultant inferred data.

Figure 4 shows the profile of a fact, of confidence 0.3, which has a 75% relevance from a 90% reliable source.

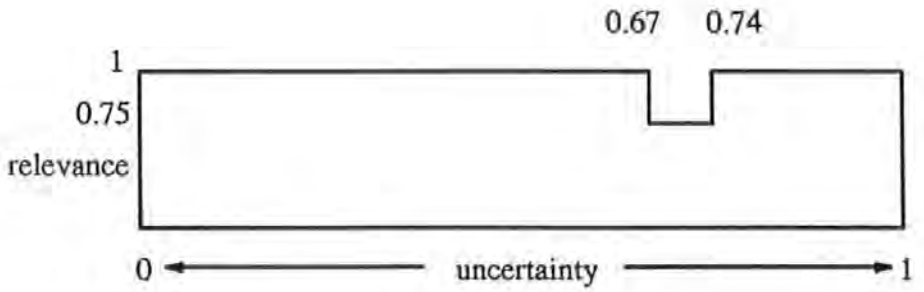


Figure 4

7.3 Rules (Intrinsic to the expert system)

Rules also require three degrees-of-integrity:
heuristic content - shown on the x-axis;
reliability - as with Facts, shown on the x-axis;
relevance - as with Facts, shown on the y-axis.

heuristic content of a rule is directly analogous to confidence with regard to facts and data. A rough and ready heuristic rule may often seem to give the correct answer but without the rigour of mathematical proof can we be sure that it applies in many

cases. The heuristic content must be reflected in the overall inferred data uncertainty profile, since the attributes of the rule are part of the expert system and invisible to the user. Rules of a high heuristic content have right orientated windows and rules that are, in fact, recognized laws have left orientated windows.

reliability is a measure of the system programmers' assessment of their confidence in the source of the rules, ie are they experts in the field? Reliability is directly analogous to its counterpart with facts: the x–axis or tolerance of the window being used.

relevance is directly analogous to its counterpart with Facts; the y–axis or depth–of–cut into the window being used.

7.4 Inferred Data (Intrinsic to the expert system)

The degrees–of–integrity for Inferred Data are:
orientation – shown on the x–axis;
intensity – shown on the y–axis;
profile – shown on the x–axis.

The process of generating the uncertainty profile of the inferred data involves superimposing the window of the rule over the windows of its associated data and facts. The resulting inferred data window being generated by a numerical integration Simpson's Rule process. The effect can be visualised as the shadows produced after light has passed through all the windows concerned. The inferred window profile is finally normalised to a thickness of unity to remove any accumulated unwanted or irrelevant data which would distort the profile.

Figures 5, 6 and 7 show inferred–data profiles for differing input data and facts.

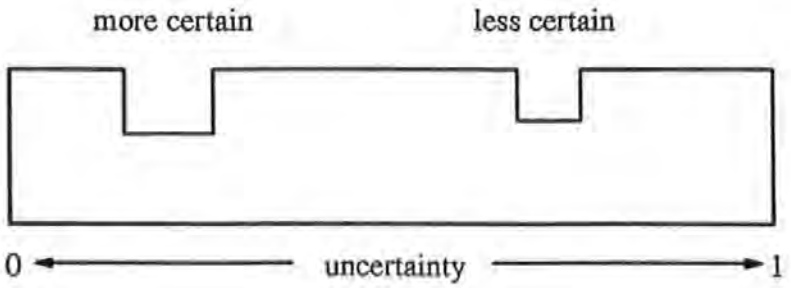


Figure 5

orientation
left orientation implies more certain inferred data;
right orientation implies less certain inferred data.

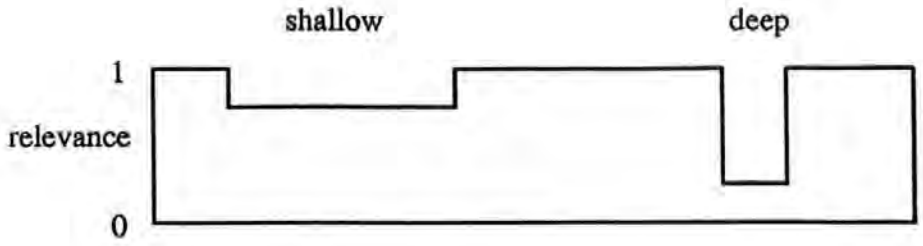


Figure 6

intensity
deep cuts imply non-relevant inferred data;
shallow cuts imply very-relevant inferred data.

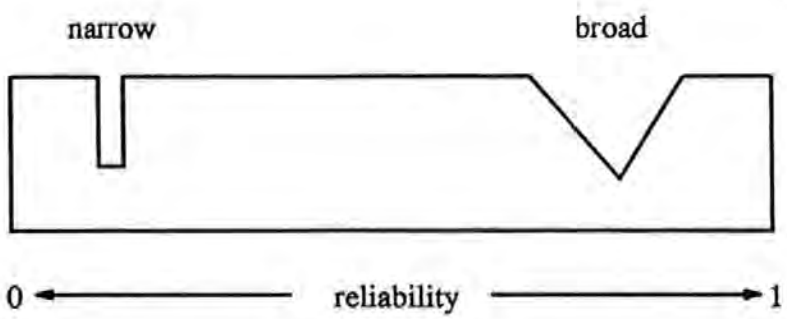


Figure 7

profile
narrow, implies reliable inferred data;
broad, implies unreliable inferred data.

8. The model in operation

In order to demonstrate the principle of the model in operation, a simple twenty heuristic rule system to select the primary, secondary and tertiary master synchronisation switch centres in a digital network is used.

8.1 Network Synchronisation

Pulse code modulation (pcm) transmission and all derived functions, such as digital switching, must be synchronised for correct operation. Individual pcm transmission links are kept in synchronisation within themselves by means of frame alignment. Here a demultiplexer extracts information (a frame alignment bit code pattern) from the incoming bit stream to derive an identical clock frequency at the receiver as at the transmitter.

In a digital network of switching nodes and transmission links a complex arrangement is necessary to maintain complete synchronisation. The switching rates

within a switching node may, typically, be governed by the frequency of a local quartz crystal oscillator of a Timing Unit (TU). This is a reliable frequency source, but there will be minor differences between the natural frequencies of the TUs at the many switching nodes in the network, since the natural frequency of the oscillations will drift from nominal with age and operating conditions. Thus the frequencies within a network of oscillators will change unless the network is locked (synchronised) to a single frequency source that has an acceptable margin of accuracy and reliability: the Primary Master oscillator. This primary master must be backed-up with a Secondary and a Tertiary Master for system reliability.

If the switch node clocks were not synchronised, the information rate of a signal received at a node would differ from the rate at which the node could process this information and possibly retransmit it. This would result in information being lost (if the input rate was faster than the receiving switch node local clock) or repeated (if it was slower).

Within such a network there are various considerations that must be incorporated into the design. These include situations where the phase and frequency of the received clock vary in a cyclic fashion called **wander**; where the physical path length of the transmission link changes due to thermal effects; and where links or nodes fail, or are broken, and so are unavailable to pass on the synchronisation signals along routes comprising a number of connected nodes and links.

The failure rate of a line system is length dependent, so the links contributing to synchronisation should be as short as possible. Links should also follow physically diverse routes as much as possible, to reduce the effect that a single fault could have on a designed network. Each link contributing to the synchronisation system requires a synchronisation unit (SU) to compare the phase of the local clock with that of the received clock for the far end of the link. Each SU should be fail-safe to ensure that its malfunction has no adverse effects. A number of links participate in the synchronisation control at each switch and thus failure of an individual transmission link or SU should not cause the clock at that switch to lose synchronisation with clocks at other switches.

The size of networks, traffic routing and complexity of control usually necessitates a hierarchical arrangement of switches for a economical solution.

Hierarchical formalisation:
level 4 = international digital exchange and primary master;
level 3 = all transit switches and the secondary master;
level 2 = all tandem switches and the tertiary master;
level 1 = all local switches.

In the event of a network comprising only local switches then the Hierarchical formalisation is modified to the following:
level 4 = public digital exchange and primary master;
level 3 = secondary master;

level 2 = all local private exchanges (PBXs) interconnected in mesh arrangement and tertiary master;
level 1 = all local PBXs interconnected in a star arrangement.

8.2 Synchronisation Cost

In developing the rule base the concept of relative cost is introduced. Differing technologies, link length and switches are allocated a cost which ranges from zero for the ideal to infinite for the unsuitable. The ideal solution to the design process is that with the minimum cost. Some simplified examples of the cost function in the expert system, with an indication of the cost principles and the constraining features are as follows.

The cost of a link is a weighted function of the traffic capacity of the link, the larger the capacity the greater the effect of failure.

The length of the link is proportional to the probability of its failure.

The cost of a switch is directly related to the number of digital links connected to it; as the failure of the switch affects its neighborhood switches via the attached links.

The relative traffic cost is taken to be zero if more than thirty erlangs, otherwise 0.5, where one erlang is defined as one circuit occupied for one hour (the standard unit for traffic measurement).

The relative distance cost is taken to be zero for distances up to 10km, 0.5 for distances between 10 and 1000km and unity for distances over 1000km.

The relative wander cost is 0 if < 1000km radio;
0 if < 1000km buried cable;
0 if < 100km aerial fibre;
0 if < 10km aerial copper;
0.5 if >= 1000km radio;
0.5 if >= 1000km buried cable;
0.5 if >= 100km aerial fibre;
0.5 if >= 10km aerial copper;
1 if >= 1500km buried cable;
1 if >= 300km aerial fibre;
1 if >= 60km aerial copper;
1 if satellite connection.

The relative switch cost is zero if the number of connected links is unity, 0.5 for between 1 and 10 links and unity for more than 10 links.

8.3 Rule Base

Rule 1: The primary master should be located at the 'centre' of the network, where the centre is a function of distance, traffic, link and switch technologies.

Rule 2: The centre of a network occurs when the maximum calculated route cost is at a minimum.

Rule 3: The maximum route cost is the sum of the link and switch costs taking the longest route between two switches in a network.

Rule 4: The link cost comprises the sum of traffic capacity, wander resistance and distance costs.

Rule 5: The cost of a switch is a function of the number of digital routes.

Using rules 1 to 5 it is possible to identify one or more switches that are network centres. If only one is found, it is necessarily the centre, otherwise the choice between them is determined by the greatest number of digital links and account is taken of its level in the hierarchy.

Rule 6: When two or more switches have the same lowest maximum route cost the switch with the greatest number of digital routes is at the centre of the network.

Rule 7: When two or more switches have the same lowest maximum route cost and the same number of digital routes, the switch which is higher in the hierarchy is the centre of the network.

Rule 8: When rule 7 has failed to select the centre, the sum of the distances of every connected link to each switch is calculated. The network centre is now the switch having the minimum distance cost.

Rule 9: When rule 8 fails to select a centre, the technology of the links is considered and the centre taken as the switch having the minimum wander cost on the links.

Rule 10: If in rule 9 there is a mixture of technologies in a single route, the worst case is taken as representing the whole distance.

By this stage, the primary master location will have been found. It is then necessary to discover the secondary and tertiary masters by repeating the above process. However at this stage, it is not necessary to find the optimum centre but rather the next best two switch locations that are closest to the primary.

Rule 11: Identify the best two neighbor switches to the primary master switch and search each, reusing rules 1 to 10, replacing the master switch cost value by zero.

The interconnection of switches must be made to a set of rules based on the hierarchy. In addition a number of resilience requirements must be satisfied.

Rule 12: A switch can only give synchronisation to another switch at the same level or at a lower level in the hierarchy. It is also necessary to prevent looping, thus the direction of control is important. Under normal operating conditions the primary must have precedence over the secondary and the secondary over the tertiary.

Rule 13: Control between masters will be unilateral from the primary to the secondary to the tertiary.

Rule 14: Control between switches at differing levels in the hierarchy is unilateral from the highest to the lowest.

Rule 15: Control between switches at the same level is bilateral.

The requirement for resilience dictates the following link requirements.

Rule 16: Every level 3 switch should be connected to a level 4 switch via no less than two synchronisation links in tandem and the precedence allocated to the links is greater to the link from the higher level than those at the same level.

Rule 17: In the event of more than one link from the same level, a higher precedence is given to the link with the smallest number of tandem links to the primary master.

Rule 18: Level 2 switches should have a minimum of two synchronisation links with precedence given to the link from the higher level.

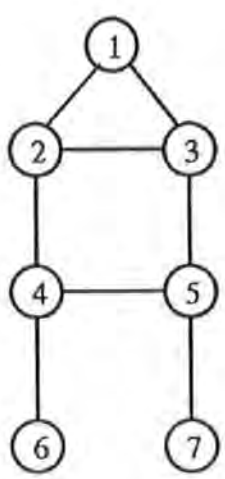
Rule 19: Level 1 switches should receive their synchronisation from a switch at level 2. If this is not possible then a minimum of two links from the same level are preferred.

Rule 20: Where no suitable primary rate link is available, a 64 kbit link to the highest level with which the switch has a traffic community of interest is required.

9. The rules in operation

In the following basic example link costs are set to unity and node costs to zero.

Example Network



Start Node	By rule 3 cost	By rule 3 route	By rule 6 digital links	By rule 7 hierarchy
1	5	1-2-3-5-4-6	2	1
2	5	2-1-3-5-4-6	3	2
3	5	3-1-2-4-5-7	3	2
4	5	4-2-1-3-5-7	3	3
5	5	5-3-1-2-4-6	3	3
6	6	6-4-2-1-3-5-7	1	1
7	6	7-5-3-1-2-4-6	1	4

Table 1

Notice that more than one switch node (1 to 5) has the same minimum cost and that conflict resolution will be necessary to select the network centre. Table 1 shows that rule 6 is used to reduce the number of centres from five to four, ie nodes 2 to 5 inclusive. By rule 7 nodes 4 and 5 may be eliminated, reducing the number of centres to nodes 2 and 3 only. Rule 8 may now be used by considering the distances involved, but now with uncertain data, as shown in table 2.

User defined distance	User's confidence	By rule 8 Involved node
1 - 2 ➸ 1	0.9	2
1 - 3 ➸ 1	0.9	3
2 - 3 ➸ 1	0.9	2 and 3
2 - 4 ➸ 2	0.8	3
3 - 5 ➸ 2	0.7	3
4 - 5 ➸ 1	0.7	
4 - 6 ➸ 3	0.5	
5 - 7 ➸ 3	0.5	

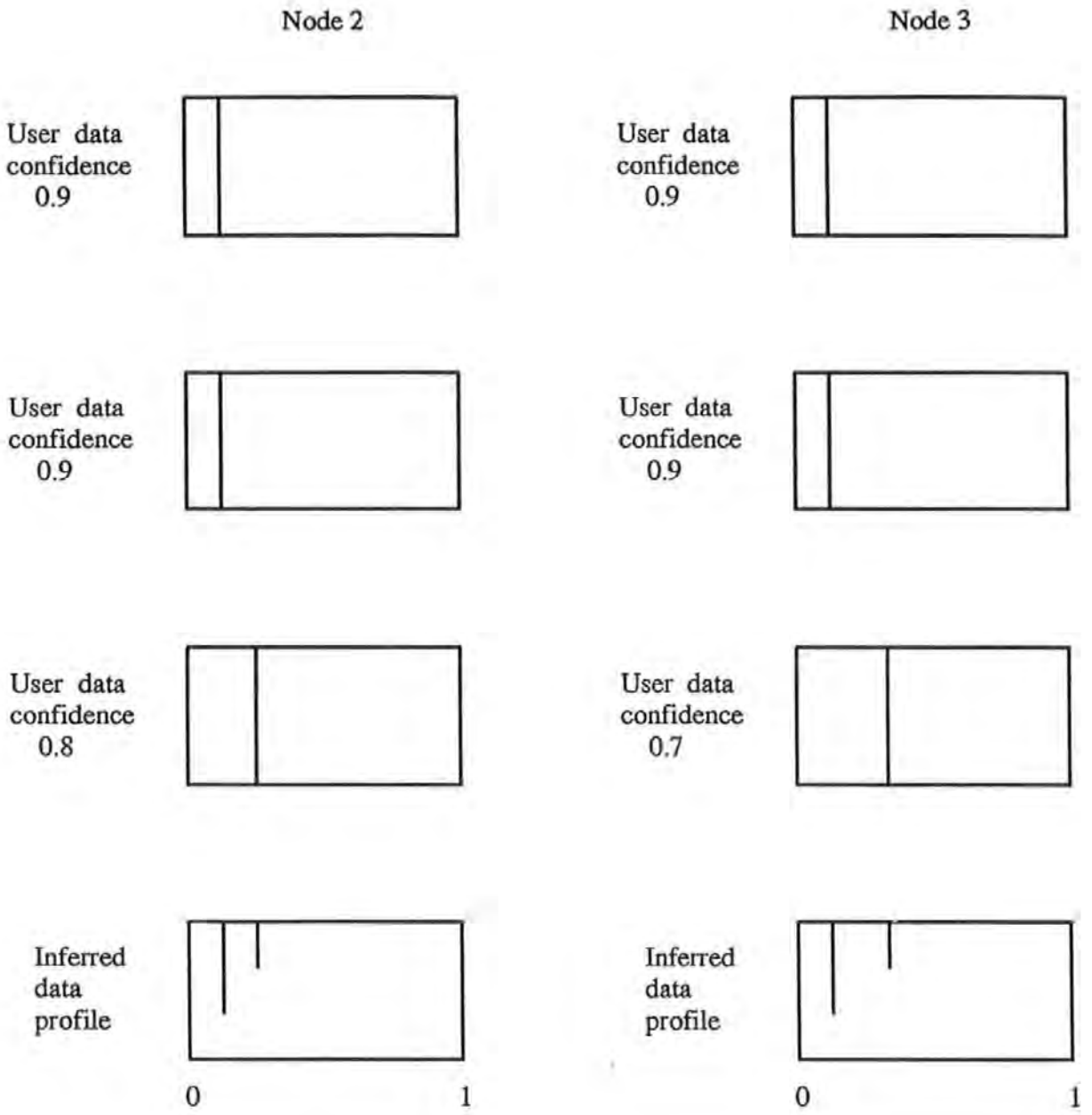
Table 2

Traditional methods for dealing-with-uncertainty would give:
node 2-distance sum 4, confidence 0.648;
node 3-distance sum 4, confidence 0.567.

The 'windows' method, shown in figure 8, shows that the original confidences attached to the data are not now lost but are reflected in the confidence profile for the inferred-data, with the orientational and intensity degrees of integrity of the inferred data being clearly shown. The actual inferred data intensities in this case are determined from the individual resultant intensities, (1+1+0)/3=0.67 and (0+0+1)/3=0.33. However, we are still unable to select the correct centre from rule 8 since both distances equal 4 and the confidence profile is approximately the same in both cases. Rules 9 and 10 are are therefore used in consideration of the technology, as shown in table 3.

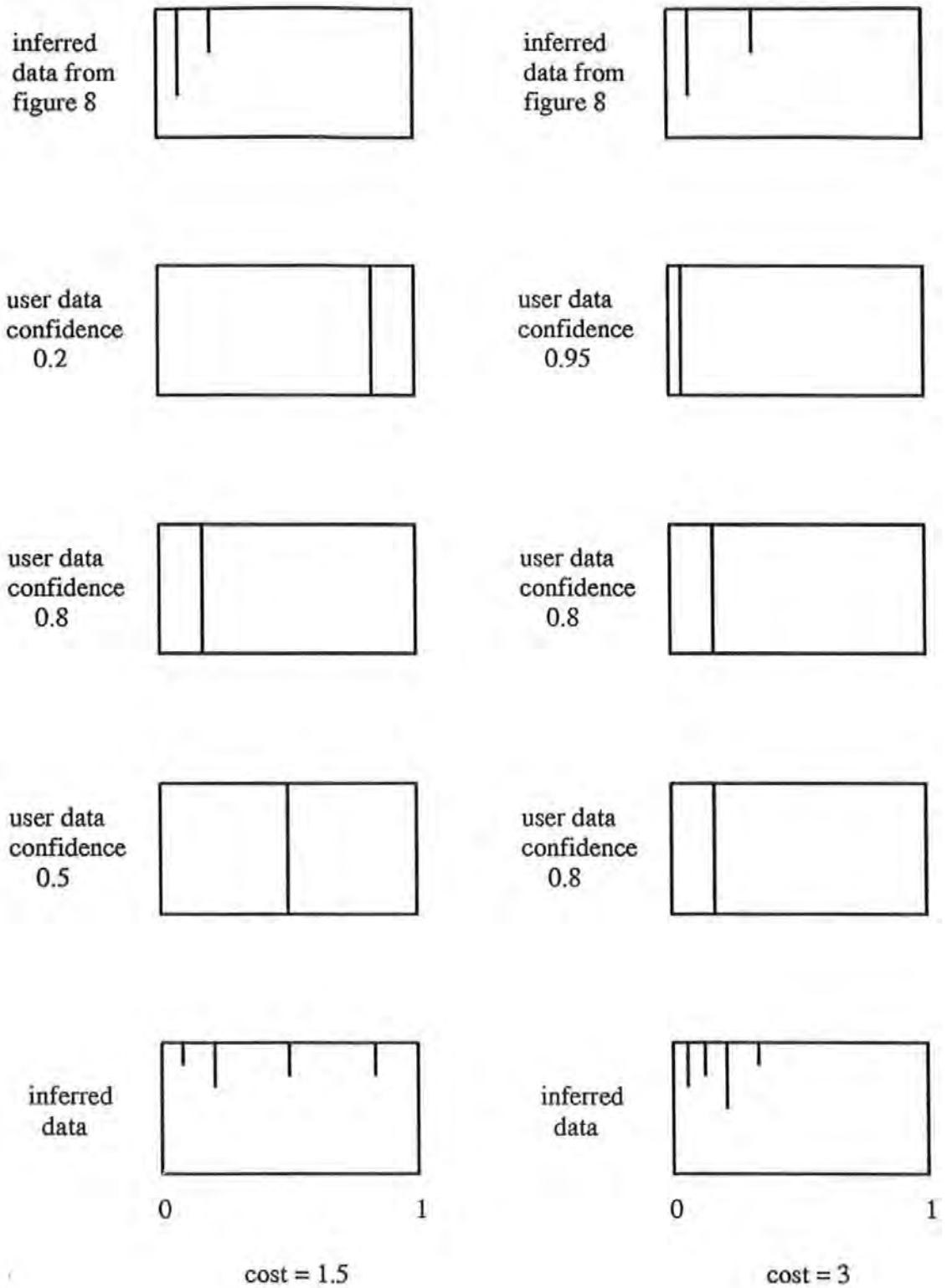
Links	User defined technology	User's Confidence	By rule 8 Involved node	By rule 9 Cost
1 - 2	900 km radio	0.2	2	0
1 - 3	900 km aerial fiber	0.95	3	1
2 - 3	1000 km aerial fiber	0.8	2 and 3	1
2 - 4	2000 km radio	0.5	2	0.5
2 - 5	2000 km fiber	0.8	3	1

Table 3



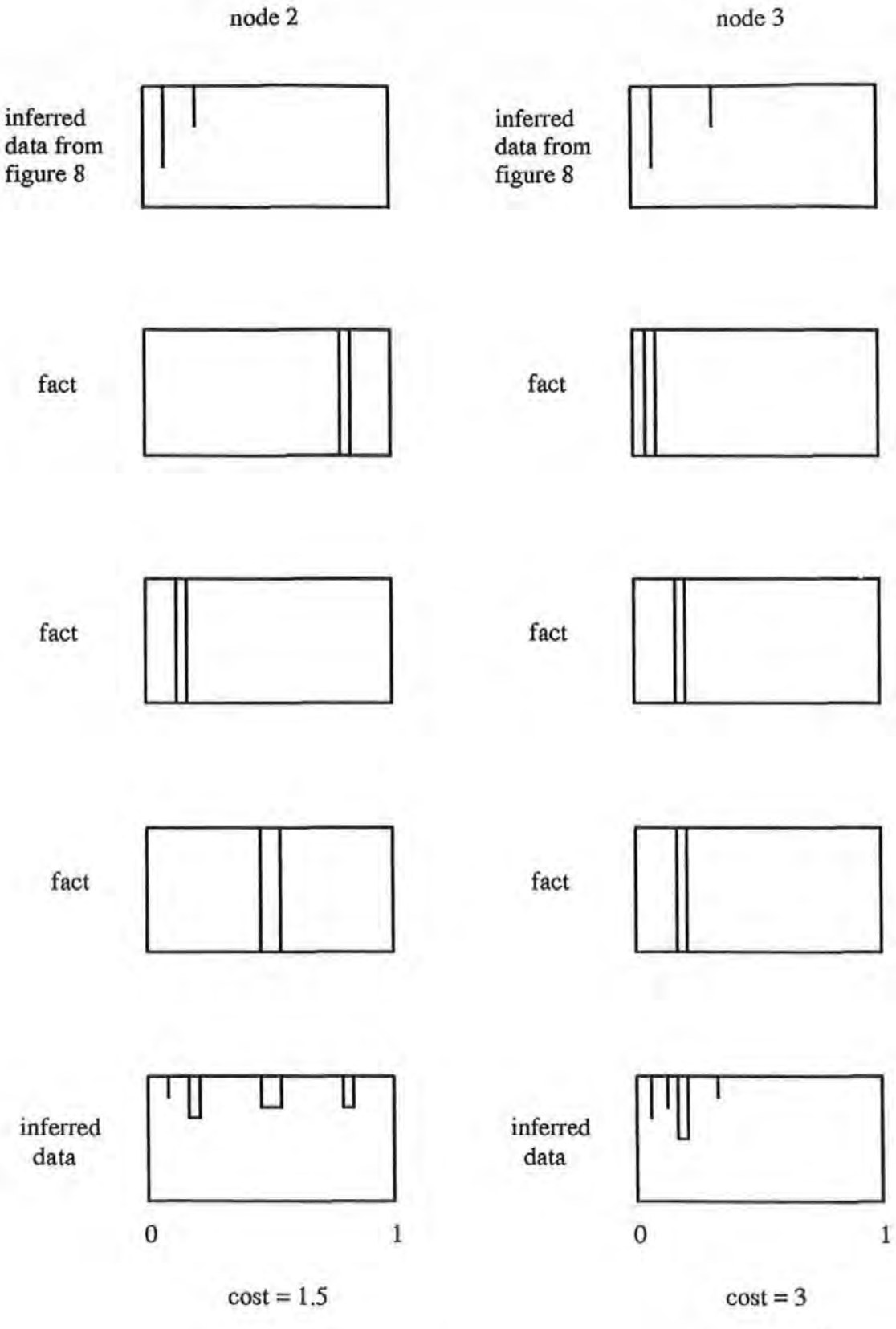
In each diagram, horizontal scale: uncertainty; vertical scale: relevance.

Figure 8



In each diagram, horizontal scale: uncertainty; vertical scale: relevance.

Figure 9



In each diagram, horizontal scale: uncertainty; vertical scale: relevance.

The shaded areas representing the reliability of the facts are shown as 5% for demonstration purposes.

Figure 10

Traditional methods for the dealing with uncertainty would give:
node 2 – sum of costs 1.5, confidence 0.08;
node 3 – sum of costs 3, confidence 0.608;
thus indicating node 2 as the network centre with its lower cost. However, how should we deal with the associated low certainty of 0.08?

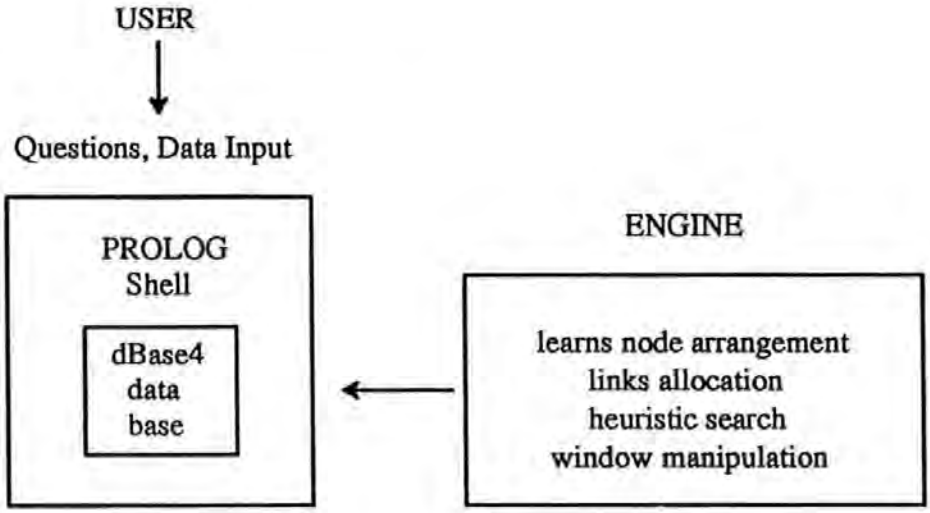
The windows method, shown in figure 9, indicates that a large proportion of the data supporting the choice of node 2 has confidence value of 0.5 and over, which suggests that the calculated confidence value of 0.08 is misplaced. There is enough high confidence data in the conclusion for node 2 to support its claim to be the network centre on the grounds of lowest cost.

Figure 10 shows the result of adding the reliability degrees of integrity. Here the spread of integrity of facts is associated with the original inferred data and the resultant new inferred data intensities determined as in the previous figures. The resultant inferred data with its associated three degrees of integrity for orientation, intensity and profile supports the selection of Node 2 as both relevant and reliable facts give a left orientated, shallow and narrow profile. This example is rather simplistic and in practice there is normally overlapping and dispersion over part of the confidence range. Even so, there is often a concentration of confidence, or it can be obtained by producing more appropriate reliable data and facts to help with an indication of the best choice for the specific application.

10. Implementation of models in Expert Systems

An Expert System shell has been developed, that demonstrates the window technique. It is written in PROLOG and dbase4, running on an IBM 386 personal computer. The construction of the shell is shown in figure 11.

One principal objective of the program to date has been to focus upon the maintenance of commonality of input by users, as in conventional expert systems. Thus **User data** is input as a 'datum and confidence' pair and Facts are input by the system designer. He is able to input the more demanding 'degrees–of–integrity' relating to the fact:
uncertainty is the experts' own confidence in the fact; ie has he used it before, is it widely used?, is it fundamental? etc.
reliability is the system designers' confidence in the expert. Does the evidence come from a text book or colleague or wife or student, etc.? This also allows the user to express his confidence in the expert.
relevance is problem specific and has a value added at run–time whilst a problem is being solved. For example, at each run of the expert system the relevance of the data increases as it is accessed by the system, such that un–accessed data can be removed at a later stage in a 'garbage collection' operation as irrelevant.



Data Base

Node data records	Number	Technology	Links	Hierarchy		
Link data records	Number	Technology	Connected to	A	B	
Routing data records	Route	A end	B end	Links	Link technology	

Window profiles	Data	0 0.001	0.0011 0.002	0.0021 0.003	0.0031 0.004	0.0041 0.005	0.0051 0.006
	Fact	0 0.001	0.0011 0.002	0.0021 0.003	0.0031 0.004	0.0041 0.005	0.0051 0.006
	Rule	0 0.001	0.0011 0.002	0.0021 0.003	0.0031 0.004	0.0041 0.005	0.0051 0.006

Figure 11

Rules are input by the system designer. He is able, once again, to input two degrees-of-integrity relating to the rule. Heuristic content reflects, if in the opinion of the expert, if the rule is of thumb, theory or scientific law; with reliability being treated as with facts. The relevance of the rule to the application is dictated by the Expert System meta-rules, which allocate a value at run-time.

Output is under the control of the **Prolog** Engine, which manipulates the data stored in dBase, conducts the heuristic searches and "window" control.

11. Conclusions

The window technique, developing in part, has been shown to be useful in the expert system for network design. In particular, it gives the user more information on which to base his decision: for example, it highlights areas of weakness which could be re-analysed to gain greater confidence in the solution produced by the expert system. The technique could be suitable for a wide range of expert systems that require users to identify problem areas with data and confidence. To date the implementation has shown reasonable results from simple to complex synchronisation networks. However, it is only one part of the network design process. Other areas, including practical problems of technology, commercial pressures and political constraints are being investigated to validate the process.

12. References

Cartes-Rello, E. and Golshai, F., 1990, Uncertainty reasoning using the Dempster-Shafer method, an application in forecasting and marketing management, Expert Systems, **17**, 9-18

Castillo, E. and Alvarez, E., 1990, Uncertainty methods in expert systems, Micro Civil Eng., **5**, 43-58

Clark, D. A., 1990, Numerical and symbolic approaches to uncertainty management in AI, Art. Intell. Rev., **4**, 109-146

Gaag, L. C., 1990, Different notions of uncertainty, Int. J. Man-machine Studies, **33**, 595-606

Garbolino, P., 1987, Bayesian Theory and AI, Int. J. Man-machine Studies, **27**, 729-742

Grout, V. M., 1988, Optimisation techniques for telecommunication networks, PhD Thesis, CNAA, University of Plymouth

Guan, J. et al., 1990, Dempster-Shafer theory, IEE Colloquium on Reasoning under Uncertainty, Digest 86, 6-13

Henkind, S. J. and Harrison, M. C., 1988, An analysis of four uncertainty calculii, IEEE Trans. Sys. Man Cybern., **118**, 700-714

Hughes, C., 1989, The representation of uncertainty in medical expert systems, Med. Info., **14**, 269-279

Ibrekk, H. and Morgan, M. G., 1987, Graphical communication of uncertain quantities to non-technical people, Risk Analysis, **7**, 519-529

Kamimura, K. et al., 1989, Expert systems supporting private network design, NTT R&D (Japan), **38**, 1377-1386

Lees, C., 1989, Defining expert systems, Telecommunications, **23**, 65-71

Magill, W. G. W. and Leech, S. A., 1991, Uncertainty techniques in expert systems software, DECIS Support Sys., **7**, 55–65

Murphy, S. V. et al., 1990, Fuzzy sets and typicality theory,
Info. Sci., **51**, 61–93

O'Leary, D. E. O. and Kandelin, N. A., 1988, Validating the weights in expert systems: a statistical approach, Int. J. Expert Sys., **3**, 252–279

Salasoo, A., 1991, Initiating usable methods with a new engineering design tool, SIGCHI Bulletin, **23**, 68–70

Sheridon, F. J. K., 1991, A survey of techniques for inference under uncertainty, Art. Intell. Rev., **15**, 89–119

Shiraishi, N. et al., 1989, Knowledge based systems for damage assessment, Proc. AI Civil and Structural Engineers, Civil Com Press, Sept. 1989, 211–216

Solo, A. et al., 1990, Setell: the strategy expert for telecommunications investment, IEEE Expert, **15**, 14–22

Spiegelhalter, D. J., 1989, Probabalistic reasoning in expert systems, J. Math. Man. Sci., **9**, 191–210

Van Der Gaag, L. C., 1990, Different notions of uncertainty,
Int. J. Man–machine Studies, **33**, 595–606

Zwick, R. and Wallsten, T. S., 1989, Combining stochastic uncertainty and linguistics inexactness, Int. J. Man–machine Studies, **30**, 69–111

EPILOGUE

It's the uncertainty in life,
That makes it very real.
It creates the emotions,
That human beings feel.
Excitement, terror,
Pain and pleasure,
Anticipating love.
If things were very certain,
There'd be none of the above.

(Lorry Reynolds - 1990)

The gods did not reveal, from the beginning,
All things to us; but in the course of time,
Through seeking, men find that which is better.
Let us conjecture that this is like truth.
But as for certain truth, no man has known it,
Nor will he know it; neither of the gods,
Nor yet of all the things of which I speak.
And even if by chance he were to utter
The final truth, he would himself not know it;
FOR ALL IS BUT A WOVEN WEB OF GUESSES.

(Xenophanes 540-500 BC)