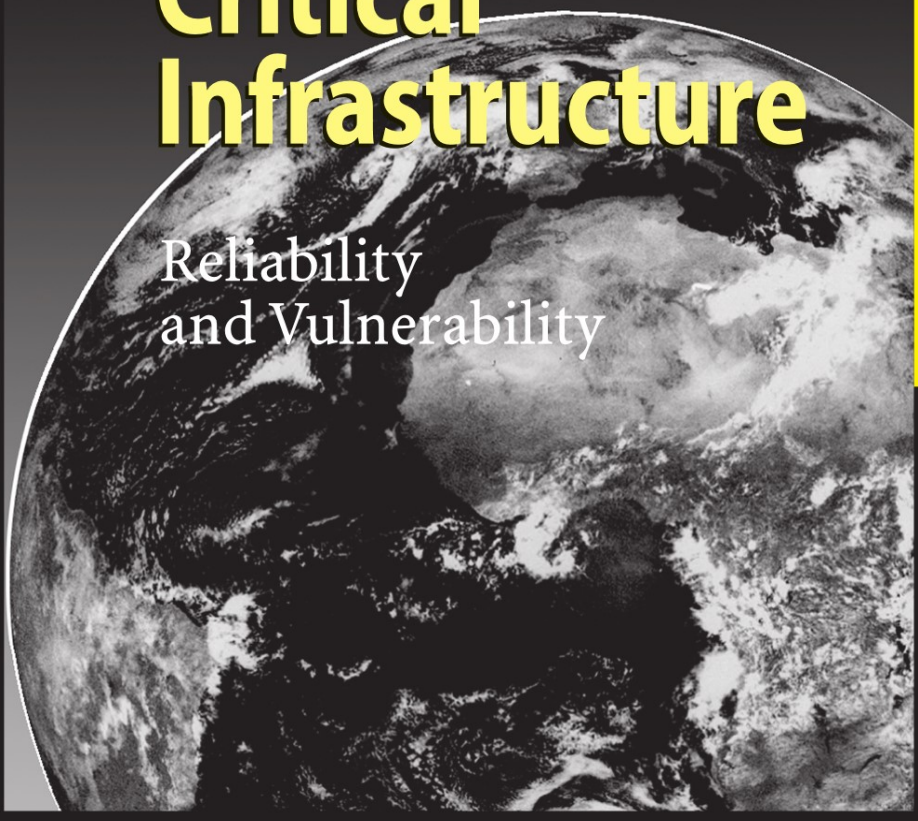
A. T. Murray
T. H. Grubesic
Editors

Critical Infrastructure

Reliability and Vulnerability



ADVANCES IN SPATIAL SCIENCE

Springer

Advances in Spatial Science

Titles in the Series

G. Clarke and *M. Madden* (Eds.)
Regional Science in Business
VIII, 363 pages. 2001. ISBN 978-3-540-41780-4

M. M. Fischer and *Y. Leung* (Eds.)
GeoComputational Modelling
XII, 279 pages. 2001. ISBN 978-3-540-41968-6

M. M. Fischer and *J. Fröhlich* (Eds.)
Knowledge, Complexity and Innovation Systems
XII, 477 pages. 2001. ISBN 978-3-540-41969-3

M. M. Fischer, J. Revilla Diez and *F. Snickars*
Metropolitan Innovation Systems
VIII, 270 pages. 2001. ISBN 978-3-540-41967-9

L. Lundqvist and *L.-G. Mattsson* (Eds.)
National Transport Models
VIII, 202 pages. 2002. ISBN 978-3-540-42426-0

J. R. Cuadrado-Roura and *M. Parellada* (Eds.)
Regional Convergence in the European Union
VIII, 368 pages. 2002. ISBN 978-3-540-43242-5

G. J. D. Hewings, M. Sonis and *D. Boyce* (Eds.)
Trade, Networks and Hierarchies
XI, 467 pages. 2002. ISBN 978-3-540-43087-2

G. Atalik and *M. M. Fischer* (Eds.)
Regional Development Reconsidered
X, 220 pages. 2002. ISBN 978-3-540-43610-2

Z. J. Acs, H. L. F. de Groot and *P. Nijkamp* (Eds.)
The Emergence of the Knowledge Economy
VII, 388 pages. 2002. ISBN 978-3-540-43722-2

R. J. Stimson, R. R. Stough and *B. H. Roberts*
Regional Economic Development
X, 397 pages. 2002. ISBN 978-3-540-43731-4

S. Geertman and *J. Stillwell* (Eds.)
Planning Support Systems in Practice
XII, 578 pages. 2003. ISBN 978-3-540-43719-2

B. Fingleton (Ed.)
European Regional Growth
VIII, 435 pages. 2003. ISBN 978-3-540-00366-3

T. Puu
Mathematical Location and Land Use Theory, 2nd Edition
X, 362 pages. 2003. ISBN 978-3-540-00931-3

J. Bröcker, D. Dohse and *R. Soltwedel* (Eds.)
Innovation Clusters
and Interregional Competition
VIII, 409 pages. 2003. ISBN 978-3-540-00999-3

D. A. Griffith
Spatial Autocorrelation and Spatial Filtering
XIV, 247 pages. 2003. ISBN 978-3-540-00932-0

J. R. Roy
Spatial Interaction Modelling
X, 239 pages. 2004. ISBN 978-3-540-20528-9

M. Beuthe, V. Himanen, A. Reggiani
and *L. Zamparini* (Eds.)
Transport Developments
and Innovations in an Evolving World
XIV, 346 pages. 2004. ISBN 978-3-540-00961-0

Y. Okuyama and *S. E. Chang* (Eds.)
Modeling Spatial
and Economic Impacts of Disasters
X, 323 pages. 2004. ISBN 978-3-540-21449-6

L. Anselin, R.J.G.M. Florax and *S. J. Rey*
Advances in Spatial Econometrics
XXII, 513 pages. 2004. ISBN 978-3-540-43729-1

R.J.G.M. Florax and *D. A. Plane* (Eds.)
Fifty Years of Regional Science
VIII, 400 pages. 2004. ISBN 978-3-540-22361-0

D. Felsenstein and *B.A. Portnov* (Eds.)
Regional Disparities in Small Countries
VI, 333 pages. 2005. ISBN 978-3-540-24303-8

A. Reggiani and *L.A. Schintler* (Eds.)
Methods and Models in Transport
and Telecommunications
XIII, 364 pages. 2005. ISBN 978-3-540-25859-9

H.W. Richardson and C.-H.C. Bae (Eds.)
Globalization and Urban Development
X, 321 pages. 2005. ISBN 978-3-540-22362-7

G. Arbia
Spatial Econometrics
XVII, 207 pages. 2006. ISBN 978-3-540-32304-4

B. Johansson, C. Karlsson,
R. Stough (Eds.)
The Emerging Digital Economy
X, 352 pages. 2006. ISBN 978-3-540-34487-2

H. Westlund
Social Capital in the Knowledge Economy
X, 212 pages. 2006. ISBN 978-3-540-35364-5

A.E. Anderssson, L. Pettersson,
U. Strömquist (Eds.)
European Metropolitan Housing Markets
approx. 380 pages. 2007.
ISBN 978-3-540-69891-3

Alan T. Murray · Tony H. Grubesic
(Editors)

Critical Infrastructure

Reliability and Vulnerability

With 66 Figures
and 51 Tables

Springer

Professor Alan T. Murray

The Ohio State University
Department of Geography
1036 Derby Hall
154 North Oval Mall
Columbus
Ohio 43210
USA
murray.308@osu.edu

Professor Tony H. Grubesic

Indiana University
Department of Geography
701 Kirkwood Ave
Student Building 120
Bloomington
Indiana 47405-7100
USA
tgrubesi@indiana.edu

Library of Congress Control Number: 2007921531

ISSN 1430-9602

ISBN 978-3-540-68055-0 Springer Berlin Heidelberg New York

Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover-design: WMX Design GmbH, Heidelberg

SPIN 11945567 88/3100YL - 5 4 3 2 1 0 Printed on acid-free paper

Preface

The aim of this text was to bring together differing geographic perspectives in modeling and analysis designed to highlight infrastructure weaknesses or plan for their protection. This began initially as the outgrowth of a series of lectures we organized for the Regional Science Association International conference in Seattle, Washington on November 10-13, 2004, and expanded substantially beyond this to include researchers in other disciplines and other countries. We are pleased with the final product and greatly appreciate the efforts of the contributors, referees, and the publisher.

January 2007

Alan Murray
Tony Grubesic

Table of Contents

1 Overview of Reliability and Vulnerability in Critical Infrastructure

Alan T. Murray[1], Tony H. Grubesic[2]

[1] Department of Geography, The Ohio State University, USA, Email: murray.308@osu.edu
[2] Department of Geography, Indiana University, USA, Email: tgrubesi@indiana.edu

1.1 Introduction

The concept of *interconnection* is an important one for a wide range of social, economic and political issues. Broadly defined, interconnection refers to a state of reciprocal connection. In this context, two or more interconnected entities can exchange ideas, currency, information and other valuable goods with each other, often for mutual benefit. For example, telecommunication backbone providers frequently interconnect at points of presence (POPs) or Internet exchanges (IXs) in order to accommodate peering relationships between large networks or to provide data transit for smaller systems. One obvious benefit accrued through this type of practice is extending the geographic reach of each backbone involved with the interconnection arrangement, providing access to new markets and potential customers. Over time, these interconnections can strengthen or decline, depending on the benefits acquired through interconnectivity. If the relationship between entities strengthens significantly, the condition of interdependency can emerge. In this context, the entities involved require the reliable operation of their interconnected partner(s) to function properly. If the relationship between entities weakens significantly, connections may be disbanded.

In recent years, the interdependencies of many infrastructure systems have increased dramatically. In the White House (2003) report titled "National Strategy for the Physical Protection of Critical Infrastructures and

Key Assets" problems associated with increased levels of interconnectivity between critical infrastructure systems are noted:

"the challenges and uncertainties presented by critical nodes and single-points-of-failure within infrastructures, as well as increasing interdependencies that exist among the various infrastructure sectors both nationally and internationally… are often difficult to identify and resolve, as are the cascading and cross-sector effects associated with their disruption" (White House 2003, pp. 33)

Perhaps the most notable problem with these increased levels of interdependencies is the potential for cascading failure across mutually dependent systems. Cascading failure occurs when an event triggering a collapse produces a series of secondary failures in interdependent infrastructures (Carreras et al. 2002; Little 2002; Albert et al., 2004; Talukdar et al. 2003; Houck et al. 2004). For example, if the electrical grid is significantly disrupted, it is likely that telecommunication services will also be disrupted. This, in fact, occurred during the massive electrical blackout in the Northeastern U.S. and portions of Canada in 2003 (Grubesic and Murray, 2006). Nearly 2,500 telecommunication networks were disrupted (*ibid*). Not surprisingly, with each disruption to a critical infrastructure system, accidental or otherwise, attempts are made to reevaluate the degree to which many of these engineered systems are able to maintain some type of operational continuity. The results of these evaluations are often used to fortify or protect existing systems, plan and construct newer, more resilient infrastructure, motivate new public policies regarding critical infrastructure and its expected performance and to help in the development of hazard mitigation plans.

1.2 Critical Infrastructure

Societal functions are highly dependent on networked systems in the developed world. Even the most basic day-to-day functions involve interaction with a variety of critical infrastructure systems. For example, millions of Americans utilize transportation infrastructure to get to work, school, or the local mall. Telecommunication infrastructure is used to maintain contact with family and friends, shop or perform financial transactions. Energy infrastructure is used to heat our homes, power local industries and deliver fuel to our automobiles. While these basic activities are relatively easy to comprehend, the magnitude of infrastructure use is less obvious. For instance, over 19 billion tons of freight valued at $13 trillion dollars was moved through the multimodal transportation system and its associated networks in the United States during 2002 (USDOT, 2006). Where

telecommunication networks are concerned, U.S. backbone traffic exceeded 100 petabytes per month in 2002 (SVBJ, 2002). Assuming an average email is 25 kilobytes, this translates into 45,035,996,273 emails per month. Finally, the daily delivery capacity of the U.S. natural gas grid is 119 billion cubic feet, with yearly consumption estimated at 22.8 trillion cubic feet during 2002. Considering the degree to which industrialized societies are reliant on such critical infrastructure systems, their importance should not be underestimated. Moreover, because the operability of these systems can be vulnerable to disasters, accidents and intentional harm, there is a need to understand how critical infrastructure and its functionality might be impacted when subjected to disruption. Thus, there is a need to develop strategies for planning networked systems capable of surviving and performing under duress.

As a response to the growing threat of terrorism in the late 1990s, the U.S. federal government established the President's Commission on Critical Infrastructure Protection (E.O. 13010). This executive order defined "infrastructure" as (E.O. 13010):

> The framework of interdependent networks and systems comprising identifiable industries, institutions (including people and procedures), and distribution capabilities that provide a reliable flow of products and services essential to the defense and economic security of the United States, the smooth functioning of government at all levels, and society as a whole.

More importantly, E.O. 13010 (1996) suggests that "…certain national infrastructures are so vital that their incapacity or destruction would have a debilitating impact on the defense or economic security of the United States." The concept of "vital" or "critical" infrastructure is important for establishing national security benchmarks. Basic inventories of critical infrastructure are often subdivided into sectors, and include (E.O. 13010, 1996; White House, 2003):

- telecommunications
- electrical power systems
- gas and oil storage
- transportation
- banking and finance
- water supply systems
- emergency services (including medical, police, fire, and rescue)
- continuity of government.

Similarly, a group of key assets were also highlighted:

- National Monuments and Icons
- Nuclear Power Plants

- Dams
- Government Facilities
- Key Commercial Assets

In this context, critical infrastructure, both in the United States and abroad, encompasses a vast array of engineered systems and assets. While no single system or asset is more important than the other, the interdependent nature of their functionality is clearly of concern. As mentioned previously, if a single system is disrupted, there is the potential for secondary failures in interdependent infrastructures. As a result, there is a significant need to both measure and monitor the reliability and potential vulnerabilities of these infrastructure systems. Further, the ability to model the effects of infrastructure failure is an important aspect of network design, disaster recovery planning, critical infrastructure identification, and fortification. Given the massive presence of economic, transportation, telecommunication, energy and medical networks in the industrialized world, it is important to have a spectrum of techniques capable of identifying potential vulnerabilities in singular network elements, or more generalized systematic weaknesses to be protected or fortified.

1.3 Reliability and Vulnerability

The concepts of reliability and vulnerability are especially important when examining the ability of critical infrastructure to provide continuity in operation. Broadly defined, reliability refers to the probability that a given element in a critical infrastructure system is functional at any given time. That is, reliability is a probabilistic measure of elements in a critical infrastructure system and their ability to *not* fail or malfunction, given a series of established benchmarks or performance guidelines. For example, it is not uncommon to assign simple reliability metrics to components in a telecommunications system.

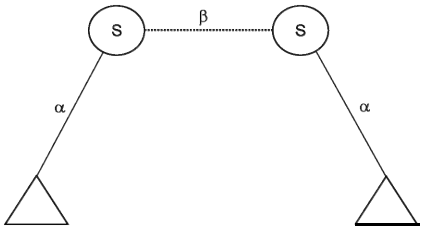


Fig. 1.1. Example telephone switching network

As an example, consider the telephone switching network given in Figure 1. Each distribution segment (α) connects a telephone to a local switch (S). In both cases, A and S have 99.99% reliability. In turn, the switches are connected to each other by a facility linkage (β), which is 99.97% reliable. The resulting reliability of this simple telephone system is 99.93% or 368 minutes of downtime per year (Medhi, 1999; Grubesic et al., 2003).

In contrast to reliability, vulnerability is a more wide-ranging concept, with much broader implications. While reliability focuses on the possibility of maintaining the performance of critical infrastructure elements, vulnerability focuses on the potential for disrupting these elements or degrading them to a point where performance is diminished. This is a subtle, yet important difference. For example, if we reexamine the simple telecommunication network in Figure 1, the concept of vulnerability becomes clearer. Let us assume that one of the distribution segments (α) is located in an area that is undergoing major flooding. Under normal operating conditions, this distribution segment is 99.99% reliable. The fact that severe weather is occurring nearby does not change this simple measure of reliability - the segment remains 99.99% reliable. However, it does make the distribution segment *more* vulnerable to disruption. For instance, Sprint/Nextel, the fourth largest U.S. wireless and wireline carrier, recently suffered two fiber cuts to its network (Reardon, 2005). As work crews performed emergency maintenance on a fiber link near the California/Nevada border that had washed out due to heavy rains, traffic from the link was rerouted to Phoenix, Arizona over a secondary path. However, this secondary link, located between Palm Springs and Phoenix, also suffered a fiber cut, disrupting long distance service to both the Sprint Nextel wireless network and the residential network along with data traffic on its backbone system (Reardon, 2005). Although the majority of customer complaints emanated from the Western U.S., service disruptions were noted nationwide.

In this context, *vulnerable* may not mean unreliable, and unreliable does not necessarily mean vulnerable. However, both reliability and vulnerability are important to the continuity of critical infrastructure operations. That said, critical infrastructure systems are exposed to a myriad of operational threats, each with a unique ability to disrupt operations. For example, road networks are vulnerable to flooding, landslides, traffic congestion, and major accidents. Telecommunication networks are vulnerable to denial of service attacks, computer viruses, targeted infrastructure attacks, and congestion. In fact, all critical network infrastructures, to some degree, are vulnerable to either technological or natural hazards. A common theme in the analysis and evaluation of network-based critical infra-

structure is *interdiction*, where network elements (nodes or links) are disabled, intentionally or otherwise, disrupting the flow of valuable goods or services through the network. Again, this could be the result of a targeted attack, accident or natural disaster. The question is, how does one identify and evaluate the degree of vulnerability for critical infrastructure elements and their systems? More importantly, once these measures of vulnerability are established, how does one take steps to minimize risk and develop infrastructure systems resistant to disruption?

1.4 Quantitative Geographic Perspective

This book focuses on techniques and approaches which utilize a quantitative geographic perspective for examining infrastructure reliability and vulnerability. It is important to note that geography, by nature, is an integrative discipline, drawing from a wide variety of quantitative techniques for solving problems with a spatial component. In fact, this integrative approach is well represented in this book, with contributions from geography, regional science, engineering, public policy, operations research, business, mathematics, economics, safety and city and regional planning.

The actual approaches outlined in this book represent a unique array of spatial analytical methods, from location modeling to network simulation, examining problems associated with the operational continuity of critical infrastructure in different ways.

For example, both simulation and optimization-based techniques have played a significant role in examining potential interdiction impacts, recognizing the insights they can provide for mitigating facility loss and prioritizing fortification efforts.

Simulation has been an important optimization technique in general terms, but certainly has proven valuable in the analysis of vulnerabilities in critical network infrastructure. One benefit of simulation is that it typically allows for the examination of a range of impacts, with either implicit or explicit notions of optimized performance for a network. In the context of evaluating the reliability and vulnerability of networks, as nodes or links are interdicted, the corresponding changes in network connectivity or performance can be documented.

One can also utilize spatial optimization approaches and network analysis for assessing impacts relative to altering the maximum flow or shortest path for a given origin-destination pair. For example, one strategy for network interdiction is to maximize network disruption by removing the links with the greatest value to a system. Similarly, one can also seek

to maximize network disruption by removing the nodes most critical to system operation. An alternative approach characterizes impacts in terms of network element attributes. For example, one can examine nodal interdiction outcomes quantified as the total attributes (e.g. capacity) of arcs impacted. Finally, one can also look at system performance by considering average service costs and coverage reduction using median and covering location models, respectively. Or, in the context of reliability, use median-based approaches for exploring the tradeoffs associated with choosing facility locations to minimize costs while taking into account expected transportation costs after failures of facilities.

Outside of simulation and spatial optimization, spatial econometric approaches can be used to estimate the economic impact of infrastructure disruption and its subsequent recovery and reconstruction activities. Not surprisingly, the "ties that bind" virtually all of these outlined approaches are geographic information systems (GIS). GIS allows for capture, management and analysis of critical infrastructure data, as well as the resulting output for visualization purposes. Recognized widely as both a tool (GIS) and a developing field of study that addresses the production of geographic data, the transformation of data into useful geographic information, and the construction of geographic knowledge (GIScience), all of the chapters included in this book utilize GIS and the basic tenets of GIScience in some way.

1.5 Summary

The motivation for this book stems from a series of special sessions organized for the North American Regional Science Association International meeting held in Seattle, Washington in November of 2004. Regional Science is an international community of scholars that has a long tradition of examining the regional impacts of national or global processes of economic and social change. A significant strength of Regional Science is its ability to draw on a wide variety of disciplines to help facilitate new theoretical and methodological insight into regional problems. As a result, this book reflects the multidisciplinary character of Regional Science and includes contributions from scholars in geography, economics, mathematics, public policy, engineering, operations research / management science, urban and regional planning, transportation, business, safety and defense.

References

Albert, R., H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature*. 406, 378-382.

Carreras, B. A., V. E. Lynch, I. Dobson, and D. E. Newman. 2002. Critical points and transitions in an electric power transmission model for cascading failure blackouts. *Chaos*. 12(4), 985-994.

E.O.1310. 1996. Critical Infrastructure Protection. URL: Executive Order 13010—*Critical Infrastructure Protection*. Federal Register, July 17, 1996. 61(138), 37347-37350. Reference is on page 37347.

Grubesic, T.H. and A.T. Murray. 2006. Vital Nodes, Interconnected Infrastructures and the Geographies of Network Survivability. *Annals of the Association of American Geographers*. 96(1), 64-83.

Grubesic, T.H., M.E. O'Kelly, and A.T. Murray. 2003. A Geographic Perspective on Commercial Internet Survivability. *Telematics and Informatics*. 20, 51-69.

Houck, D. J., E. Kim, G. P. O'Reilly, D. D. Picklesimer, and H. Uzunalioglu. 2004. A network survivability model for critical national infrastructures. *Bell Labs Technical Journal*. 8(4), 153-172.

Little, R. G. 2002. Controlling Cascading Failure: Understanding the Vulnerabilities of Interconnected Infrastructure. *Journal of Urban Technology*. 9(1), 109-123.

Medhi, D., 1999. Network reliability and fault-tolerance. In: Webster, J. (Ed.), Wiley Encyclopedia of Electrical and Electronics Engineering. John Wiley and Sons, New York.

Reardon, M. 2005. Spring Nextel suffers service outage. *CNET News*. http://www.news.com. January 9.

SVBJ [Sillicon Valley Business Journal]. 2002. U.S. Internet Traffic Tops 100 Petabytes. URL: http://sanjose.bizjournals.com/sanjose/stories/2002/05/06/daily40.html

Talukdar, S. N., J. Apt, M. Ilic, L. B. Lave, and M. G. Morgan. 2003. Cascading Failures: Survival versus Prevention. *The Electricity Journal*. November, 25-31.

USDOT [United States Department of Transportation]. 2006. Freight in America. URL: http://www.bts.dot.gov/publications/freight_in_america/

White House. 2003. The National Strategy for the Physical Protection of Critical Infrastructures and Key Assets. URL: http://www.whitehouse.gov/pcipb/physical_strategy.pdf

2 Transport Network Vulnerability: a Method for Diagnosis of Critical Locations in Transport Infrastructure Systems

Michael A. P. Taylor, Glen M. D'Este

Transport Systems Centre, University of South Australia, Australia;
Email: map.taylor@unisa.edu.au

2.1 Introduction

Considerations of critical infrastructure are now a major concern in Australia as in many other countries. The concern stems from a variety of causes, including the state of development, condition and level of use of existing infrastructure systems, especially transport networks; difficulties associated with public sector provision of new infrastructure; public-private partnership arrangements for infrastructure provision; and perceptions of risks and threats to infrastructure from both natural disasters (e.g. floods, fire or earthquake) and from human malevolence such as acts of sabotage, war or terrorism. The Australian Federal Government has defined critical infrastructure as 'that infrastructure which if destroyed, degraded or rendered unavailable for an extended period, will significantly impact on social or economic well-being or affect national security or defence' (Attorney-General's Department, 2003). A pertinent question is then how to identify critical locations in an infrastructure network. For example, the road transport network is large, wide and diverse in nature. Are there particular locations or facilities in that network where loss or degradation of certain road sections (links) will have significant impacts? How should such impacts be assessed? Thus there are needs for the develop-

ment and application of a methodology to assess risk and vulnerability of transport networks. Methods and decision support tools are needed that allow planners and policy makers to make rational assessments of threats to facilities and infrastructure; the consequences of network degradation and failure at various locations and under different circumstances; and what to do about these. Social and economic benefits flow from the ability to plan for and manage the impacts of transport network degradation to minimise wider consequences on economic, employment, trade and social activities in cities and regions.

This chapter provides an overview of our recent research on developing a methodology for transport network vulnerability analysis, based on considerations of the socio-economic impacts of network degradation. At one level this involves considerations of alternative paths through a network and the relative probabilities of use of those paths. Whilst probability of use is important in defining potential weak spots in a network, this probability is not of itself a complete measure of vulnerability – the most critical locations in a network will show the most severe (socio-economic) consequences resulting from network failure at those locations. The methods therefore consider vulnerability assessment in terms of a planning systems process in which the performance of network components is tested against established performance criteria. The risks and consequences associated with failures at different locations need to be accounted for. Suitable metrics that may be used to interpret the extent and consequence of network failure or degradation need to be developed and tested.

The concept of network vulnerability is new, and it is important to define what is meant by vulnerability. For instance, there are several possible responses to the reduced performance of a degraded network, or in dealing with the perceived risks of degradation at different locations. In some cases, an appropriate response may be to upgrade key transport infrastructure, for instance by raising it above expected maximum flood levels or by adding more capacity. But sometimes this simply makes the network more reliant on those key links and more vulnerable to their failure. An alternative approach is to add links to the network. These links may normally be redundant but provide alternative routes when key network links are broken. At the urban network level there may already be many such latent alternative routes, but at the regional or national strategic network level this is less likely to be the case. Extra links would make the transport network more robust, but this may add unnecessary cost to the provision of transport infrastructure. The question is where are these locations of potential network vulnerability and what is the best response.

The starting point for our study of network vulnerability was the study of transport network reliability, which has been the subject of intense in-

ternational research interest over the last decade, following the Kobe earthquake of 1995.

2.2 Network Reliability

Transport network reliability has the subject of considerable international research interest in recent years (Lam 1999, Bell and Cassir 2000, Iida and Bell 2003, Nicholson and Dante 2004). Much of this research has focused on congested urban road networks and the probability that a network will deliver a required standard of performance. The urban studies are important, but they are not the only areas of concern, especially when considering the wider implications of transport systems performance. At the regional and national strategic level, accessibility, regional coverage and inter-urban connectivity are the primary considerations. In these sparse networks, 'vulnerability' of the network can be more important than 'reliability' because of the potentially severe adverse consequences of network degradation. As noted by the Bureau of Transport and Resource Economics (BTRE 1999) in its analysis of the effects of flooding on road access,

> 'the vast distances involved means that access to alternative services (such as hospitals and business) often do not exist … disruption costs to households, businesses and communities can therefore be more important in rural and remote communities'.

In both urban and rural areas, the concept of vulnerability or incident audit – the proactive determination of locations in a transport network that may be most sensitive to failure and where network failure may have the gravest consequences – requires detailed research. The transport planner may seek opportunities to reduce vulnerability – and the community will demand such action.

Network reliability became an important research topic in transport planning during the 1990s, although some elements had been the subject of research interest for some time before that (e.g. Lee 1946, Richardson and Taylor 1978, Taylor 1982). The Kobe earthquake of 1995 and its aftermath stimulated an interest in *connectivity reliability*. This is the probability that a pair of nodes in a network remains connected – i.e. there continues to exist a connected path between them – when one or more links in the network have been cut. Bell and Iida (1997) provided an analytical procedure for assessing connectivity reliability, and a summary of the procedure is given by Iida (1999). Subsequent research was directed at degraded networks, usually urban road networks subject to traffic congestion, in which the network remained physically intact but the performance of one or more links could be so severely affected by congestion that their use by traffic is

curtailed. This led to the definition of two additional forms of reliability: travel time reliability and capacity reliability, as described below.

Travel time reliability considers the probability that a trip between an origin-destination pair can be completed successfully within a specified time interval (Bell and Iida 1997). This can be affected by fluctuating link flows and imperfect knowledge of drivers when making route choice decisions (Lam and Xu, 2000). One measure of link travel time variability is the coefficient of variation of the distribution of individual travel times (Richardson and Taylor, 1978). Measures of travel time variability are useful in assessing network performance in terms of service quality provided to travellers on a day-to-day basis (Yang, Lo and Tang, 2000). Thus travel time variability can be seen as a measure of demand satisfaction under congested conditions (Asakura, 1999).

A supply-side measure of network performance in congested networks is *capacity reliability* (Yang, Lo and Tang 2000). Capacity reliability is defined as the probability that a network can successfully accommodate a given level of travel demand. The network may be in its normal state or in a degraded stated (say due to incidents or road works). Chen, Lo, Yang and Tang (1999) defined this probability as equal to the probability that the reserve capacity of the network is greater than or equal to the required demand for a given capacity loss due to degradation. Yang, Lo and Tang (2000) indicated that capacity reliability and travel time reliability together could provide a valuable transport network design tool. Taylor (1999, 2000) demonstrated how the concepts of travel time reliability and capacity reliability could be used in planning and evaluating traffic management schemes in an urban area.

Further research on network reliability is required to develop these concepts into practical traffic planning tools. In addition, there is a need for further research to properly specify travellers' responses to uncertainty (Bonsall 2000, Van Zuylen 2004) so that reliability research can be used to properly inform developments of new driver information systems and to influence the design of traffic control systems.

2.3 Network Vulnerability

From the above review, we may conclude that the standard approaches to transport network reliability have focused on network connectivity and travel time and capacity reliability. While this provides valuable insights into certain aspects of network performance, reliability arguments based on probabilities and absolute connectivity may obscure potential network

problems, especially in large-scale, sparse regional or national networks. In these networks the consequences of a disruption or degradation of the network become important. For example, D'Este and Taylor (2001) used the example of the Australian land transport system to illustrate the potential consequences of the severance of certain transport connections in this multimodal network. In this example the system reliability was considered, in terms of a cut to the Eyre Highway and transcontinental rail line between Perth and Adelaide, for instance by flood. The overall network remains connected and the probability that the route in question is cut by flood or other natural cause is extremely small (but not zero since it has happened), so the travel time and capacity reliabilities are high. Therefore the established measures of network reliability would not indicate any major problem with the network. However the consequences of network failure are substantial – in this case the next best feasible path through the network involves a detour of some 5000 km. Nicholson and Dalziell (2003) pointed to similar circumstances in their study of the regional highway network in the centre of the North Island of New Zealand, a region subject to both snowstorms and volcanic eruptions.

These examples illustrate the concept of network vulnerability and the difference between network reliability and vulnerability. The concept of vulnerability is more strongly related to the consequences of link failure, irrespective of the probability of failure. In some cases, link failure may be statistically unlikely but the resulting adverse social and economic impacts on the community may be sufficiently large to indicate a major problem warranting remedial action – akin to taking out an insurance policy for an extremely unlikely yet potentially catastrophic event. For example, consider the impact on a rural community of loss of access to markets for its produce and to vital human services (such as a hospital). Low probability of occurrence and network performance elsewhere does not offset the consequences of a network failure. Thus network reliability and vulnerability are related concepts but while reliability focuses on connectivity and probability, vulnerability is more closely aligned with network weakness and consequences of failure. Berdica (2002) proposed that vulnerability analysis of transport networks should be regarded as an overall framework through which different transport studies could be conducted to determine how well a transport system would perform when exposed to different kinds and intensities of disturbances. From her study of the road network in central Stockholm she suggested three main questions that might be posed in these studies:

1. How do interruptions of different critical links affect system performance, and to what extent?

2. How is network performance affected by general capacity reductions and possible changes to traffic management and road space allocation in a subregion of the network?
3. How is the system affected by variations in travel demand?

These questions provide a starting point for the development of a methodology for study of vulnerability in transport networks and infrastructure. They highlight the key issue of the identification of critical components of the networks. Vulnerability analysis is intended to address these questions and the perhaps more important questions that flow from them – when we know where the vulnerable elements (the 'weakest links') of a transport network are, what is the best response, what can we do about it?

2.3.1 Vulnerability and Risk

Vulnerability, reliability and risk are closely linked concepts. In broad terms, risk is something associated with negative outcomes for life, health, or economic or environmental condition. Risk can be defined in many different ways, but most definitions focus on two factors: the probability that an event with negative impacts will occur, and the extent and severity of the resultant consequences of that event. Commonly, the product of probability and a measure of consequence is used as an index of risk. This may be shown schematically as a 'risk matrix', as in Figure 2.1.

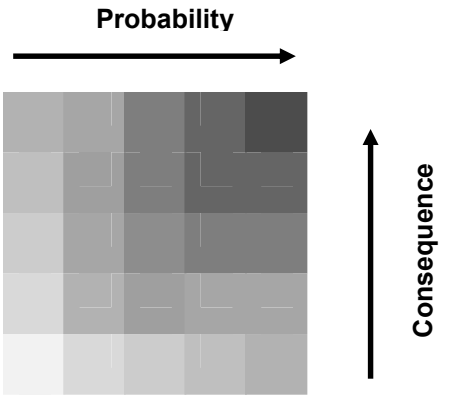


Fig.2.1 Conceptual risk matrix

Risk and reliability analysis is mostly concerned with the top-right sector of the matrix where increasing probability and increasing consequences combine. Nicholson and Dalziell (2003) applied this framework to the risk

assessment of transport networks in New Zealand. They measured risk as simply the sum of the products of the event probabilities and the economic costs of the event (e.g. the expected annual economic cost of a given event). Their risk evaluation process involved the following steps:

1. establish the context (i.e. the technical, financial, legal, social and other criteria for assessing the acceptability of risk)
2. identify the hazards (i.e. the potential causes of closure)
3. analyse the risks (i.e. identify the probabilities, consequences and expectations)
4. assess the risks (i.e. decide which risks are acceptable and which are unacceptable).

If any risk is found unacceptable, it needs to be managed. This generally involves either (1) treating the unacceptable risks, using the most cost-effective treatment options, or (2) monitoring and reviewing the risks (i.e. evaluating and revising treatments).

The study of vulnerability extends this risk assessment framework in several important ways. Firstly it extends the region of interest to areas of high consequences and low or unquantifiable (but non-zero) probability of occurrence – on the basis that measurement of occurrence probability and consequences (human and economic) is imprecise for many types of incidents, and society may well consider some consequences to be unacceptable and worthy of safeguarding against, despite uncertainty about their probability of occurrence (e.g. Evans, 1994). Secondly, vulnerability analysis provides a framework for targeting risk assessment. One of the key conclusions of the Nicholson-Dalziell risk assessment of the New Zealand highway network was that it is impractical and financially infeasible to conduct detailed geophysical and other risk assessment across an entire transport network. The costs of deriving accurate location-specific risk probabilities across a range of risk factors are too high to make it viable – what is needed is a way of targeting risk assessment resources to get best value from them. Vulnerability analysis provides another way of approaching this problem. It can be used to find structural weaknesses in the network topology that render the network vulnerable to consequences of failure or degradation. Resources can then be targeted at assessing these 'weak links'. Thirdly, vulnerability auditing admits a more proactive and targeted approach to the issue of transport network risk assessment and mitigation.

2.3.2 Definitions

The authors have defined vulnerability by using the notion of accessibility, i.e. the ease by which individuals from specific locations in a region may

participate in activities (e.g. employment, education, shopping, trade and commerce) that take place in other physical locations in and around the region and by using a transport system to gain access to those locations (Taylor and D'Este 2004a). Then vulnerability is defined in the following terms:

- a network *node* is *vulnerable* if loss (or substantial degradation) of a small number of links significantly diminishes the accessibility of the node, as measured by a standard index of accessibility
- a network *link* is *critical* if loss (or substantial degradation) of the link significantly diminishes the accessibility of the network or of particular nodes, as measured by a standard index of accessibility.

This broad definition can then be further refined by the selection of specific indices of accessibility. Amongst others, Morris, Dumble and Wigan (1979), Koenig (1980), Niemeier (1997) and Primerano (2003) provide discussions of alternative accessibility indices. For the case of strategic level networks such as a regional or national network, relatively simple indices are appropriate. A number of indices may be considered as useful in strategic network analysis. Two of these are: (1) generalised travel cost, for the elemental separation between two locations, and (2) the Hansen integral accessibility index (Hansen 1959) which provides an overall measure of the accessibility of one location to a set of other locations.

Generalised cost c_{ij} is the overall assessed cost of travel from origin i to destination j in the network. It may be taken as the network travel distance, travel time, money cost or some other measure (e.g. fuel used) between the two locations, or as a (weighted) sum of these.

These two indices are useful in assessing accessibility between major population or activity centres. In the case of regional analysis involving locations outside major population centres, some other measure of accessibility is needed. This is of particular interest in vulnerability studies of regional and remote areas such as those comprising the geographic mass of mainland Australia. Specifically for such sparsely settled regions, the Australian Government has adopted a 'remoteness' index known as ARIA (Accessibility/Remoteness Index of Australia) that is used by federal departments to assess the level of government and private sector services (e.g. in health, finance and social welfare) available to residents of regional and remote areas (DHAC, 2001). Whilst this chapter does not consider the use of ARIA, our wider research on vulnerability does, because the consequences of network degradation on rural communities in regional and remote areas are of significant societal concern. Sekhar and Taylor (2005) provides an introductory account of the study of vulnerability at the regional level.

Taking generalised cost, we can formulate a basic model that may be used to provide a measure of vulnerability in terms of the change in generalised cost of travel between two locations if a given link fails, where the generalised cost may be taken as an appropriate measure of disutility of travel such as distance, time, money, etc – in other words, the loss of amenity from link failure. Generalised cost is seen as a simple measure of elemental accessibility as it indicates the difficulty involved in travelling between the two locations (if not the overall impact of that difficulty). Let Δc_{ij}^{rs} denote the change in generalised cost of travel from node *i* to node *j* if network link e_{rs} fails (e_{rs} is a link connecting nodes *r* and *s* in the network). Then the loss of community amenity is $d_{ij}\Delta c_{ij}^{rs}$ where d_{ij} is the demand for movement from *i* to *j*. It follows that the total loss of amenity from the failure of e_{rs} is then

$$V_{rs} = \sum_i \sum_j d_{ij} \Delta c_{ij}^{rs} \tag{1}$$

The two measures, Δc_{ij}^{rs} and V_{rs}, provide local and global measures of the consequences of failure of link e_{rs}. Hence they are direct measures of the extent to which the operation of the transport system is vulnerable to failure of specific links. Note that similar definitions can be developed for node failures.

In more formal terms, the problem can be stated as follows. Consider a network *G(N,E)* where *N* is a set of *n* nodes and *E* is a set of *m* directed links. Associated with each link is a non-negative attribute that measures the utility of the link according to a particular link characteristic, such as distance, time, money cost, reliability, or generalised cost. Let s[ij, *G(N,E)*] be the 'cost' of the least cost path from origin *i* to destination *j* in *G(N,E)* then

$$\Delta c_{ij}^{rs} = \text{s[ij, } G(N,E- e_{rs})\text{]} - \text{s[ij, } G(N,E)\text{]} \tag{2}$$

that is, the difference between the least cost path with the network intact and the least cost path without the link from *r* to *s*, e_{rs}. The task of calculating Δc_{ij}^{rs} and V_{rs} can be tackled by a number of approaches, as described in D'Este and Taylor (2003) and Taylor and D'Este (2004a, b), and the essence of these approaches is the concept of a 'network scan' in which the first step is to identify candidate critical links, either because they form part of a minimum cost path between origin node *i* and destination node *j*, or in a multipath model have a reasonable probability of use for travel be-

tween those two nodes, and the next step is to fail each of those candidate links in turn and assess the consequences of those failures. Such scans can be used with any of the three indices of accessibility cited above, and relative changes in accessibility for the degraded networks then used to assess vulnerability.

2.3.3 A Specific Accessibility Index

The Hansen integral accessibility index (A_i) for location (city) i is written as

$$A_i = \sum_j B_j f(c_{ij}) \tag{3}$$

where B_j is the is the attractiveness of location (city) j, e.g. the number of opportunities available at j. In the strategic network application described in this chapter B_j is taken as the population of city j. Equation (2) is often used in a normalised form, viz

$$A_i = \frac{\sum_j B_j f(c_{ij})}{\sum_j B_j} \tag{4}$$

and this is the version used in our research, where the Hansen index has been used to consider changes in accessibility between the Australian mainland capital cities (Adelaide, Brisbane, Canberra, Darwin, Melbourne, Sydney and Perth) for degradations of the Australian National Highway System (NHS) road network. The NHS is shown in Figure 2.2.

The impedance function $f(c_{ij})$ of equations (2) and (3) represents the separation between the two cities and is defined so that the higher the cost of travel between the two cities, the lower the accessibility between them. The definition adopted in this current work is the conventional reciprocal of travel distance in the network.

2.4 The Australian Road Network

The analysis reported in this chapter is based on the NHS road network, which forms the basic skeleton of the national road system of Australia (see Figure 2.2). This subset of the national main road network has been designated by the Australian federal government as of prime importance in providing a national road transport system, and the funding for the provi-

sion and maintenance of the NHS is the express responsibility of the federal government. The full main road network connecting cities and regions is of course much more extensive than the NHS network (see Figure 2.3). The full main road network may be split into three subnetworks, which relate to the national, state or regional importance of the individual roads and highways. Besides the NHS, the other subnetworks are the state highways and designated main roads, which provide connectivity at the state level and are the direct responsibility of the state governments, and the other main roads, which provide regional connectivity and for which responsibility may be shared between state and local government. Figure 2.3 highlights the NHS and state highways and designated roads subnetworks as a skeleton amidst the matrix of the full network. In the more densely settled regions of the south east, east coast and south west, there is a substantial main road network. The network coverage away from those regions, in the less settled parts of the nation, is much sparser and here the NHS and state highways really do represent almost the entirety of the navigable road system. This may be seen in Figure 2.4, which shows the NHS and the designated state highway networks.

Fig. 2.2 The Australian National Highway System (NHS) network, connecting the major cities

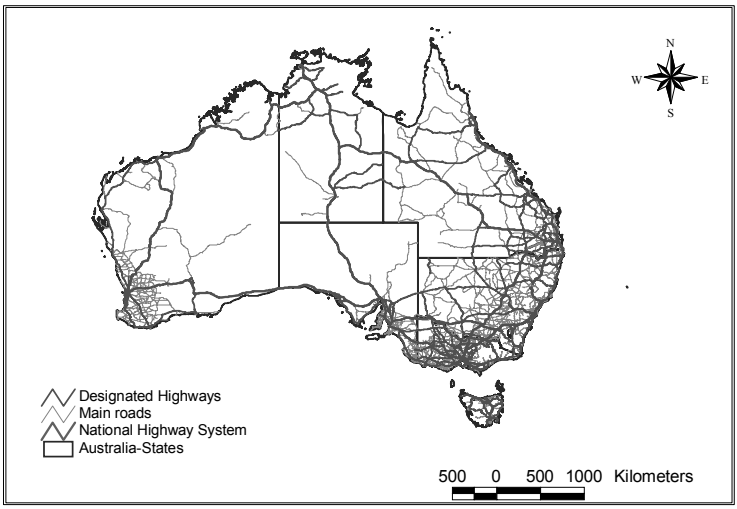


Fig. 2.3 The Australian main road network showing all main roads, designated state highways and the National Highway System (NHS) subnetworks

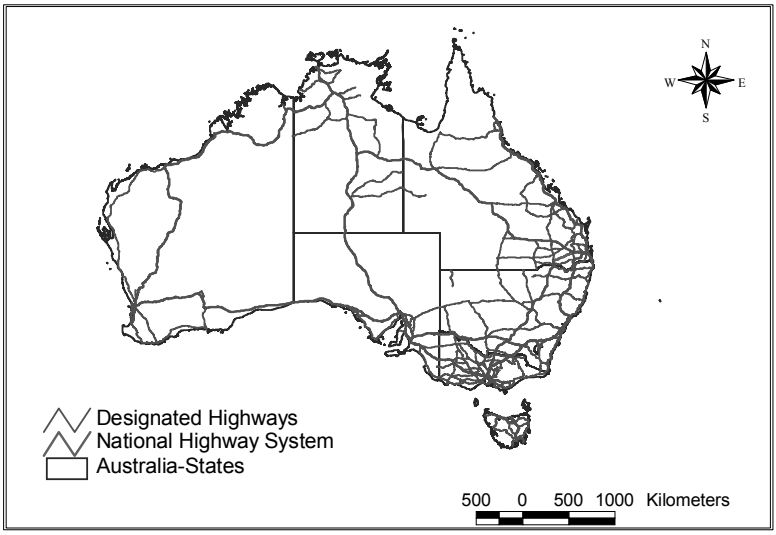


Fig. 2.4 The Australian NHS and designated state highways form a subnetwork of the full Australian main road network

A GIS database of the entire strategic road network of Figure 2.3 has been set up using the ArcGIS package. This database holds a number of attributes for all of the identified road links, including:

- road classification (NHS, stage highway, other main road)
- road type (e.g. freeway, divided carriageway, two-lane two-way road)
- region (urban, regional, remote)
- pavement type (sealed or unsealed)
- speed limit
- average operating speed
- bridge locations

In addition, attributes concerning pavement condition and traffic volume (AADT) are being progressively added to the database as they become available, using data supplied by the various state road authorities.

This full database will be used to study vulnerability at national, state and regional levels and to locate critical locations (links and nodes) in the network, using the accessibility indices and the network scanning procedures discussed previously.

2.5 Sample Network Scan of the NHS

We now present an illustrative application of our vulnerability scan methods, using the NHS network (see Figure 2.2, and also Figure 2.4) as a case study. As such, this example is restricted to considerations of the accessibility provided by the NHS as the sole network for travel between the mainland capital cities. This is a simplification of the real world situation but it clearly exemplifies the techniques for network scans and vulnerability analysis and suggests a way forward for further studies of more complex networks.

As suggested previously, a simple accessibility index is that of generalised travel costs between origins and destinations. Given origin-destination flows, then overall increases in cost in a degraded network can be assessed by using equation (1). However, such flows are not always available – and this is the current case for the Australian national road transport system. Whilst there are data on corridor flows – see for instance Gargett and Sidebottom (2003) – these do not provide information on all of the specific origin-destination movements. Future research will attempt to overcome this problem. As one approximation in the absence of observed origin-destination flows, weighted travel times can be used as a measure of change in generalised cost accessibility. We do this here to provide an indication of a vulnerability analysis based on origin-destination flows, be-

cause this relates immediately to the primary definition of vulnerability provided earlier in this chapter, and expressed by equation (1). Table 2.1 shows the (2001) populations of the mainland capital cities and the travel distances between them, using the full NHS network. Table 2.2 shows the equivalent travel times between the cities.

Table 2.1 Capital city populations and inter-city travel distances in the full NHS Network

	Ade-laide	Bris-bane	Can-berra	Darwin	Mel-bourne	Perth	Sydney
Population (2001)	1002127	1508161	339727	71347	3160171	1176542	3502301
Travel distance via NHS (km)							
Adelaide	-	1985.7	1167.5	2622.9	722.5	2691.7	1341.9
Brisbane		-	1109.0	3103.0	1536.1	4643.2	873.5
Canberra			-	3756.3	636.2	3828.6	235.5
Darwin				-	3345.5	3465.8	3873.2
Melbourne					-	3414.3	810.6
Perth						-	3999.5
Sydney							-

Table 2.2 Minimum travel times in full NHS Network

Travel time (h) *from*	*To*						
	Ade	Bri	Can	Dar	Mel	Per	Syd
Adelaide	-	23.36	13.74	30.86	8.50	31.67	15.79
Brisbane		-	13.05	36.51	18.07	54.63	10.28
Canberra			-	44.19	7.49	45.04	2.77
Darwin				-	39.36	40.77	45.57
Mel-bourne					-	40.17	9.54
Perth						-	47.05
Sydney							-

A weighting factor for each origin-destination pair was devised, based on a simple gravity model for interactions between the cities, i.e. the normalised weight w_{ij} for travel between cities i and j was

$$w_{ij} = \frac{g_{ij}}{\sum_{ij} g_{ij}} \tag{5}$$

where

$$g_{ij} = \frac{B_i B_j}{x_{ij}^2} \tag{6}$$

in which B_i is the population of city i and x_{ij} is the network travel distance between i and j. The network scan approach was to find the minimum travel time paths between origin-destination pairs and then to cut the network at each link (e) of the minimum path in succession and calculate the resulting changes Δt_{ij}^e in inter-city travel times. By summing over all origin-destination pairs and using the weights w_{ij} given by equation (5), an overall weighted network travel time increase (T_e) can be calculated:

$$T_e = \sum_{ij} w_{ij} \Delta t_{ij}^e \tag{7}$$

This analysis, summarised by Table 2.3, indicated that four particular road sections were the most critical in the NHS network. These four sections were

1. Hume Freeway, between Melbourne and Seymour
2. Hume Highway, between Albury and Gundagai
3. Hume Freeway, between Yass and Sydney
4. Ipswich Motorway, between Brisbane and Ipswich.

Link closures on these four sections yielded increases in weighted average travel times at least 5.5 times larger than those for any other road sections in the network. The next road section of interest was the South Eastern Freeway from Adelaide to Melbourne, which produced an increase in weighted average travel time of 1.66 hours, compared to the 9.13 hours for the Ipswich Motorway (see Table 2.3).

Weibull (1976), cited in Morris, Dumble and Wigan (1979), provided a set of criteria for indicating the usefulness of an accessibility index. On the basis that accessibility refers to the ease of movement to or from a place, Weibull suggested that one property of a proper accessibility index would be that its value would increase as the accessibility of the place increased. Unfortunately, generalised cost fails this test. An increased in the generalised cost of travel to or from a place indicates that it has become less accessible. In addition, Morris, Dumble and Wigan distinguished between *elemental accessibility* indices, that indicate the level of accessibility between two locations (say i and j), and *integral accessibility* indices, which indicate the overall accessibility of a given location (say i) to all other locations. Integral accessibility measures are more useful in considering the overall impacts of change of network or travel conditions. Generalised cost is a measure of elemental accessibility.

Table 2.3 Increases in weighted average travel times in degraded NHS network, from each origin city to all destination cities

Increases in weighted travel times (h) with cut to Hume Highway (Albury-Gundagai)							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total
0.06	0.00	0.96	0.00	5.72	0.01	4.82	11.57
Increases in weighted travel times (h) with cut to Hume Freeway (Yass-Sydney section)							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total
0.05	0.00	0.82	0.00	4.76	0.01	5.72	11.36
Increases in weighted travel times (h) with cut to Hume Freeway (Melbourne-Seymour)							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total e
0.00	0.00	0.63	0.00	5.07	0	3.99	9.69
Increases in weighted travel times (h) with cut to Ipswich Motorway (Brisbane-Ipswich)							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total e
0.06	4.57	0.07	0.00	3.23	0.01	1.20	9.13

An integral accessibility index whose values increase as the home location (origin) becomes more accessible, such as the Hansen index of equation (3), is intuitively more appealing than the generalised cost index, especially given a lack of data on origin-destination flows, as discussed above. A network vulnerability scan using the Hansen index was performed for the mainland capital cities in the NHS network, to illustrate the use of this index in vulnerability analysis. Table 2.4 shows the computed Hansen indices for each city and for all of the cities, when the full NHS network is available. These computations were based on the populations and inter-city network travel distances given in Table 2.1.

Table 2.4 Hansen accessibility indices in full NHS Network

	Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney
Hansen index	0.0871	0.0773	0.02148	0.0294	0.0999	0.0272	0.1120
Total Hansen index summed over all cities = 0.6477 (overall accessibility metric)							

A vulnerability scan was then undertaken, similar to that performed for the weighted mean travel time accessibility index. In this scan, each link of the minimum travel time path tree from each city was broken in turn, new minimum paths determined for the degraded networks, and revised values of the Hansen indices computed for the degraded networks. Critical road sections were determined on the basis of their impacts, when broken, on

the overall accessibility of the network. Table 2.5 summarises the results of this analysis. The accessibility metrics in this table are relative Hansen indices, i.e. the Hansen index for the degraded network divided by the relevant Hansen index for the full network. The same four road sections were identified as critical (most vulnerable) parts of the network, in terms of the reduced levels of overall accessibility (between all cities) for the NHS network. A closure in the Sydney-Yass section of the Hume Freeway leads to an overall reduction in the total accessibility of all the capital cities of some 34 per cent, and this was the most critical section of the network identified in the analysis (see Table 2.5).

It should be noted that the overall effect of the Ipswich Motorway is due to the peculiar topology of the NHS network and the simplification of any analysis of the road system based on this network alone. The Ipswich Motorway is the only road link in the NHS connecting Brisbane (Australia's third largest city) and all of the other capital cities. Closure of this road *in this network* thus has a catastrophic impact on travel to or from Brisbane, bringing it to a complete stop! In the real world network there are alternative routes that would be used – the effect of a closure of the motorway would still be important, but not as complete as in this simplified illustrative example.

Table 2.5 Relative values of Hansen accessibility index in degraded NHS network, as proportions of index values for full network

Proportionate Hansen accessibility index with cut to Hume Freeway (Sydney-Yass section)							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total
0.842	0.989	0.390	1.000	0.631	0.912	0.697	0.662
Proportionate Hansen accessibility index with cut to Hume Highway (Albury-Gundagai section)							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total
0.822	1.000	0.806	0.998	0.578	0.901	0.617	0.780
Proportionate Hansen accessibility index with cut to Ipswich Motorway							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total
0.911	0.000	0.939	0.846	0.871	0.875	0.788	0.785
Proportionate Hansen accessibility index with cut to Hume Freeway (Melbourne-Seymour section)							
Adelaide	Brisbane	Canberra	Darwin	Melbourne	Perth	Sydney	Total
1.000	0.876	0.853	1.000	0.552	1.000	0.726	0.818

A further advantage of the Hansen accessibility index is that it also explicitly reveals the effects on the individual cities of link closures, as can be seen in Table 2.5. For example, a closure on the Sydney-Yass section of the Hume Freeway leads to a 61 per cent decrease in the accessibility of the national capital Canberra, a 16 per cent decrease in the accessibility of Adelaide, an a nine per cent decrease in the accessibility of Perth[1]. Likewise, the Ipswich Motorway has a 100 per cent effect on the accessibility of Brisbane (in the NHS, see above) and a 21 per cent decrease in accessibility overall. These individual changes as well as the overall change help to more clearly define the vulnerability of specific road sections, whereas these results are not available for the generalised cost index. Figure 2.5 identifies these critical road sections on a map of the NHS network.

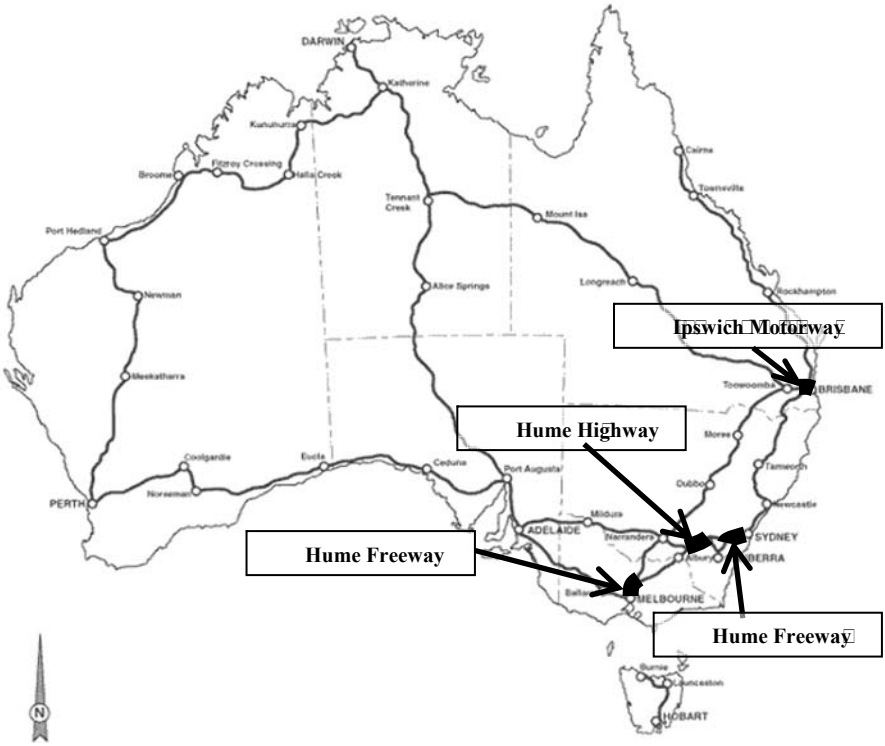


Fig. 2.5 Critical road sections from the vulnerability scan of the NHS

[1] As indicated on the previous page and shown in Table 2.5, the overall accessibility of all of the capital cities decreases by 34 per cent with a cut to the Sydney-Yass Section of the Hume Freeway.

2.6 Discussion

This chapter has discussed the development of techniques to identify specific 'weak spots' – critical infrastructure – in a network, where failure of some part of the transport infrastructure would have the most serious effects on access to specific locations and overall system performance. The Australian National Highway System road network is used as a simple case study, but the concepts and techniques described in this paper have much wider application. In particular and as a next part of our research and development of the vulnerability method, we will be adapting and applying the methods for use in the much larger and more complex road networks that exist in the real world. What we can say at present is that our research has yielded useful concepts and a method for analysis of network vulnerability in terms of the spatial or topological configuration of the network and possible socio-economic impacts assessed in terms of changes in accessibility to markets, service and facilities resulting from site-specific failure of transport infrastructure. Further research is needed to:

- develop more efficient algorithms for network vulnerability scans in large and complex networks
- develop better and more comprehensive vulnerability metrics
- refine techniques for identifying network weaknesses
- extend and refine the use of network vulnerability indicators for use in studies of critical infrastructure and the implications of network degradation
- develop techniques for recommending and evaluating cost-effective risk management and remedial responses (such as reducing risk profile, upgrading existing infrastructure, adding alternative routes, and so on). This may involve trading off the level of resources put into managing the risk against a measure of vulnerability that takes into account the implications of network failures as well as path probabilities
- develop visualisation tools for interpreting and communicating results

Candidate vulnerability metrics belong to a composite set including:

- indices of network connectivity and accessibility
- probability distributions for travel times and costs to specified destinations
- measures of change in the utility of travel
- spatial distributions of changes in the above metrics
- indices of risk, including expected values of costs, changes in these values under different network conditions, propensity for component failure, and performance thresholds.

The research results provided in this chapter suggest that an integral accessibility index such as the Hansen index can account for most of these desired properties and can be widely applied for strategic level analysis using commonly available data such as city population and network travel distances.

Our set of measures is being designed to reflect both the intensity of vulnerability and its extent, both spatially and demographically, across a study region. The techniques to apply these measures to vulnerability analysis will be based on the complex system paradigm, thus focusing the research on the required methodology, process and tools. Validation of the techniques will require careful appraisal of the modelled consequences of network failure for real world systems.

In the longer term we envisage a form of network scanning that might be termed 'incident audit' – akin to road safety audit – being developed in the long term. The aim is to provide a methodology that can identify where infrastructure failure will have the worst consequences for movement of people and goods. It includes tools for engineers and planners to determine critical network locations, and devise strategies and remedial measures to safeguard network performance. These tools can be applied at a variety of planning levels, from strategic planning to tactical planning and operational management and control.

Acknowledgement

The research reported in this paper is supported by a Discovery Grant from the Australian Research Council. The contributions of Dr Sekhar Somenahalli to the establishment of the GIS database of the Australian main road network should be acknowledged, along with the vital input of traffic and highway data by Mr David Brown of Main Roads Western Australia.

References

Asakura, Y. 1999. Evaluation of network reliability using stochastic user equilibrium. *Journal of Advanced Transportation 33* (2), pp.147-158.

Attorney-General's Department. 2003. Trusted information sharing network for critical infrastructure protection. Attorney-General's Department, Australian Government. Canberra ACT, 10 March 2003.

Bell, M.G.H.and C. Cassir (eds.). 2000. *Reliability of Transport Networks*. Baldock, Herts: Research Studies Press.

Bell, M.G.H. and Y. Iida. 1997. *Transportation Network Analysis*. Chichester: John Wiley and Sons.

Berdica, K. 2002. An introduction to road vulnerability: what has been done, is done and should be done. *Transport Policy* 9, 117-127.

Bonsall, P.W. 2000. Travellers' response to uncertainty. In *Reliability of Transport Networks*. Edited by M.G.H. Bell and C. Cassir. Baldock, Herts: Research Studies Press, 1-10.

BTRE. 2002. Benefits of flood mitigation in Australia. Report 106, Bureau of Transport and Regional Economics, Canberra.

Chen, A., Yang, H., Lo, H.K. and Tang, W.H. 1999. A capacity related reliability for transportation networks. *Journal of Advanced Transportation* 33 (2), 183-200.

D'Este, G.M. and Taylor, M.A.P. 2001. Modelling network vulnerability at the level of the national strategic transport network. *Journal of the Eastern Asia Society for Transportation Studies* 4 (2), 1-14.

D'Este, G.M. and Taylor, M.A.P. 2003. Network vulnerability: an approach to reliability analysis at the level of national strategic transport networks. In *The Network Reliability of Transport*. Edited by Y. Iida and M.G.H. Bell. Oxford: Pergamon-Elsevier, 23-44.

DHAC. 2001. Measuring remoteness: accessibility/remoteness index of Australia (ARIA) (Revised edition). Occasional Paper, New Series no 14. Commonwealth Department of Health and the Ageing, Canberra.

Evans, A.W. 1994. Evaluating public transport and road safety measures. *Accident Analysis and Prevention* 26, 411-428.

Gargett, D. and Sidebottom, A. 2003. Freight between Australian cities 1972 to 2001. Information sheet 22, Bureau of Transport and Regional Economics, Canberra.

Hansen, W.G. 1959. How accessibility shapes land use. *Journal of the American Institute of Planners* 25, 73-76.

Iida, Y. 1999. Basic concepts and future directions of road network reliability analysis. *Journal of Advanced Transportation* 33 (2), 125-134.

Iida, Y. and Bell, M.G.H. (eds.). 2003. *The Network Reliability of Transport*. Oxford: Pergamon-Elsevier.

Koenig, J.G. 1980. Indicators of urban accessibility: theory and application. *Transportation* 9, 145-172.

Lam, W.H.K. (ed). 1999. Special issue on transport network reliability. *Journal of Advanced Transportation* 33 (2).

Lam, W.H.K. and Xu, G. 2000. Calibration of traffic flow simulator for network reliability assessment. In *Reliability of Transport Networks*. Edited by M.G.H. Bell and C. Cassir. Baldock, Herts: Research Studies Press, 139-157.

Lee, C.E. 1946. New works for wartime traffic. *The Railway Magazine* 92, May/June, 177-83.

Morris, J.M., Dumble, P.L. and Wigan, M.R. 1979. Accessibility indicators for transport planning. *Transportation Research* 13A, 91-109.

Nicholson, A.J. and Dalziell, E. 2003. Risk evaluation and management: a road network reliability study. In *The Network Reliability of Transport*. Edited by Y. Iida and M.G.H. Bell. Oxford: Pergamon-Elsevier, 45-59.

Nicholson, A.J. and Dante, A. (eds.). 2004. *Proceedings of the Second International Symposium on Transportation Network Reliability (INSTR04)*. Department of Civil Engineering, University of Canterbury, Christchurch, New Zealand.

Niemeier, D. A. 1997. Accessibility: an evaluation using consumer welfare. *Transportation* 24, 377-396.

Primerano, F. 2003. Towards a policy-sensitive accessibility measure. *Papers of the Australasian Transport Research Forum* 27, paper no 35, CD-ROM. Wellington: Transit New Zealand.

Richardson, A.J. and Taylor, M.A.P. 1978. A study of travel time variability on commuter journeys. *High Speed Ground Transportation Journal* 12 (1), 77-99.

Sekhar, S.V.C. and Taylor, M.A.P. 2005. GIS approach to understanding the relationship between road network accessibility and socio-economic indicators. *Proc 4th Asia Pacific Transport and Environment Conference*. Xian, PRC. November.

Taylor, M.A.P. 1982. Travel time variability – the case of two public modes. *Transportation Science* 16 (2), 517-521.

Taylor, M.A.P. 1999. Dense network traffic models, travel time reliability and traffic management II: Application to reliability. *Journal of Advanced Transportation* 33 (2), 235-251.

Taylor, M.A.P. 2000. Using network reliability concepts for traffic calming – permeability, approachability and tortuosity – in network design. In *Reliability of Transport Networks*. Edited by M.G.H. Bell and C. Cassir. Baldock, Herts: Research Studies Press, 217-242.

Taylor, M.A.P. and D'Este, G.M. (2004a). Critical infrastructure and transport network vulnerability: developing a method for diagnosis and assessment. *Proceedings of the Second International Symposium on Transportation Network Reliability (INSTR04)*. Christchurch, August. Department of Civil Engineering, University of Canterbury, 96-102.

Taylor, M.A.P. and D'Este, G.M. 2004b. Safeguarding transport networks: assessment of network vulnerability and development of remedial measures. *Australian Journal of Multidisciplinary Engineering, Special Edition on Engineering a Secure Australia*, 13-22.

Van Zuylen, H.J. 2004. The effect of irregularity of travel times on departure time choice. *Proceedings of the Second International Symposium on Transportation Network Reliability (INSTR04)*. Christchurch, August. Department of Civil Engineering, University of Canterbury, 253-259.

Weibull, J.W. 1976. Axiomatic approach to the measurement of accessibility. *Regional Science and Urban Economics* 6, pp.357-379.

Yang, H., Lo, H.K. and Tang, W.H. 2000. Travel time versus capacity reliability of a road network. In *Reliability of Transport Networks*. Edited by M.G.H. Bell and C. Cassir. Baldock, Herts: Research Studies Press, 119-138.

3 A Framework for Vulnerability Assessment of Electric Power Systems

Åke J. Holmgren

Division of Safety Research, Royal Institute of Technology (KTH), Sweden, and the Swedish Defence Research Agency (FOI); Email: ake.holmgren@foi.se

3.1 Critical Infrastructure Protection

The infrastructure of a society consists of facilities such as communications, power supplies, transportation, water supplies, and the stock of buildings. In a broad definition of infrastructure, it is also possible to include basic societal functions like education, national defense, and financial and judicial systems. Here, the notion critical infrastructure will refer to the collection of large technical systems, for example electric power grids, which form the basis for most activities in a modern society, and are of great importance for the economic prosperity. Today, critical infrastructure protection is also considered to be a matter of national security.[1]

This chapter introduces a *framework for quantitative vulnerability assessment* (vulnerability analysis and evaluation) of critical infrastructure systems. The framework is applied to *electric power delivery* (i.e. electric power transmission and distribution). Vulnerability is described as a susceptibility (sensitivity) to threats and hazards that substantially will reduce the ability of the system to maintain its intended function.

Disturbances in the electric power supply can originate from natural disasters, adverse weather, technical failures, human errors, labor conflicts, sabotage, terrorism, and acts of war. A disturbance has its starting point in

[1] American security policy makes a distinction between "homeland security" and "national security". Critical infrastructure protection is identified as a "critical mission area" in the *National Strategy for Homeland Security* from 2002. However, protection of the infrastructures has traditionally been an integral part of the defense in countries such as Sweden and Norway (embraced by concepts such as "total defense" and "societal security").

an initiating event, i.e. a threat or hazard that is materialized. This event is, in turn, leading to one or more technical conditions in the power system that may lead to a smaller or larger power system failure and possibly a loss of electric power for all, or some, users, i.e. a power outage (black-out). In this chapter, a national security perspective is adopted, and the focus is, thus, on events that can cause severe stress on the whole society. For electric power delivery, this means power outages with a prolonged duration, a large power loss, and many affected people.

It is possible to find a broad range of checklists and practical frameworks for risk and vulnerability analysis. That is, step-by-step descriptions of how to conduct a specific method or how to use a particular analysis technique, as well as worksheets for conducting surveys (e.g. DoE 2002a; 2002b; IEC 1995). There are, however, few general frameworks that approach the subject of quantitative vulnerability assessment in a more scholarly manner. Relevant knowledge (modeling and analysis techniques) can be found in many scientific disciplines including mathematics, statistics, electric power systems engineering etc. (see also the other chapters). In order to be able to properly use quantitative techniques, there is a need for a fundamental discussion about the context of the quantitative modeling, as well as concepts such as “vulnerability” and “reliability”.

3.2 Electric Power Delivery and Major Power Outages

An electric power system can schematically be divided into generation units (generators, transformers etc.), delivery systems, and users. Where the *electric power delivery system* usually consists of:

- *Transmission grids* (high-voltage) are meshed networks, connecting large generating stations (e.g. hydro power and nuclear power), sub transmission grids, and very large users. Transmission grids enable power trading with other countries and facilitate the optimization of generation within a country.
- *Sub transmission grids*, or regional grids, are radial or locally meshed networks connected to the transmission grid via infeed points. Smaller generating plants (e.g. wind power stations and gas turbines), and large users are connected to these grids.
- *Distribution grids* (low-voltage) are radial networks that carry the electric power from the higher voltage levels to the final users. The number of levels in a distribution grid depends upon the density and magnitude of demand and the terrain.

In an electric power system there always has to be a *balance* between the load and the generation (the real time power balance stage is called dispatch). The *load* on the system varies over day and season, and so does the available *generation*. These conditions put special requirements on the operation and control of the electricity generation and delivery process. In general, there are three levels of control: i) The control center or Energy Management System (EMS); ii) The data collection system called SCADA (supervisory control and data acquisition system); iii) AGC (automatic generation control) for maintaining the instantaneous power balance.

The impact of a *major power outage* will be determined by the nature of the affected area, the duration of the disturbance, the time of day, the weather conditions etc. A major blackout will affect all functions in a society, and economical life stops in a region without electricity (UCTE 2003; U.S.-Canada Task Force 2004). People in large cities will usually be more affected than those living in rural areas. Indirect effects of a blackout can have a major spread in time and space, for example an increase in crimes in larger cities, interruptions in communications and transportations, and low indoor temperatures during wintertime. Especially critical is the state of dependence between telecommunications and power systems. After a few days there can be a shortage of food and fuel, which affects the reserve supply of electricity from backup generators.

3.3 Vulnerability Assessment

3.3.1 The Vulnerability Concept

The concept of vulnerability is employed in e.g. psychology, sociology, political science, economics, epidemiology, biology, environmental and geosciences, and engineering (McEntire 2005). For technical applications there is no generally accepted definition of the concept. In Holmgren and Molin (2005) the following working definition is used: “Vulnerability is the collection of properties of an infrastructure system that might weaken or limit its ability to maintain its intended function, or provide its intended services, when exposed to threats and hazards that originate both within and outside of the boundaries of the system”.

In this chapter, the concept of vulnerability is used to describe a lack of robustness and resilience in relation to various threats and hazards. Threats and hazards are the sources of potential harm or situations with a potential for harm. *Hazards* relate to accidental events, whereas *threats* relate to deliberate events. *Robustness* signifies that the system will retain its system structure (function) intact (remains unchanged or nearly unchanged) when

exposed to perturbations, and *resilience* implicates that the system can adapt to regain a new stable position (recover or return to, or close to, its original state) after perturbations. Here, robustness and resilience taken together is treated as the *complement of vulnerability* in the same way as safety can be an antonym to risk. (However, more refined distinctions can be made, e.g. Hansson and Helgesson (2003) show in a formal concept analysis that robustness can be treated as a special case of resilience.)

The *monadic* concept "vulnerability", divides systems into two categories: vulnerable, and not vulnerable. The *comparative* notion "at least as vulnerable as" compares systems according to their degrees of vulnerability. A monadic concept can, in theory, be obtained from the comparative one through the addition of precise limit somewhere on the scale of degrees of vulnerability. A monadic notion of vulnerability is not useful in real life – all systems are sensitive to some threats and hazards, and hence vulnerable in some respect. However, using the comparative notion is not always straightforward. A system may be vulnerable with respect to some threats (perturbations) but not to others. If two systems are vulnerable in relation to different kinds of threats, there may be no evident answer to the question which of them is more vulnerable. They may very well be incomparable in terms of vulnerability.

In this chapter, the following *formal definition of vulnerability* is proposed: the vulnerability of an infrastructure system is the probability of at least one disturbance with negative societal consequence Q larger than some large (critical) value q, during a given period of time T. Let $Q(t)$ be the societal consequence of a disturbance that occurs at time t, $t \in T$. Then, the vulnerability of the infrastructure system is measured by the function

$$P(\max_{t \in T} Q(t) > q). \tag{1}$$

Consequently, the vulnerability of an infrastructure system is the probability of a system collapse causing large negative societal consequences.

The consequence Q of a power outage can be described by technical indicators such as power loss (MW) or unserved energy (MWh). Also, more general indicators can be employed, for example the cost of the power outage or the number of affected users. No attempt will be made here to exactly specify what constitute large negative consequences (large q). However, the term severe strain on society (frequently used in Swedish official policy documents) can be used to loosely characterize what represents a major disturbance.

In some situations it is possible to estimate the probability that a hazard or threat is realized, however, in other situations (e.g. antagonistic threats),

a conditional approach can be used. Let A_i be an initiating event, then the *conditional vulnerability* can be defined as

$$P(\max_{t \in T} Q(t) > q \mid A_i). \tag{2}$$

In a study of road network vulnerability, Jenelius et al. (2006) discuss the vulnerability concept, and referrers to the conditional probability as "exposure" (See further Sect. 3.4).

There are obvious similarities to the risk and reliability concepts in the vulnerability measure above. However, the *risk* concept is both a bit more restricted and a bit broader. As with the risk concept, there are two dimensions: the probability or likelihood of a negative event and the resulting negative consequences. Risk is often reserved for random/uncertain events with negative consequences for human life and health, and the environment. Regarding the vulnerability of the critical infrastructures, planned attacks play an important role. Further, it is principally a focus on the survivability of the system, and the concept of vulnerability is not used in relation to minor disturbances. The *reliability* concept can be captured by several different measures. The reliability function (survivor function) $R(t)$ for an unreparied unit can be defined as $R(t) = P(T > t)$, for $t > 0$, where T is the time to failure. Accordingly, $R(t)$ is the probability that the unit survives the time interval $(0, t]$ (Høyland and Rausand 1994).

3.3.2 The Vulnerability Assessment Framework

A framework for quantitative vulnerability assessment of infrastructure systems is presented in Fig. 3.1. The framework draws on experiences from system studies conducted by the Swedish Defence Research Agency (FOI), and the traditional framework for risk assessment as presented in IEC (1995). Examples of vulnerability assessment frameworks inspired by the conventional risk analysis can also be found in Einarsson and Rausand (1998) and Doorman et al. (2006).

The aim of a *vulnerability assessment* can be to identify events that can lead to critical situations (large negative consequences), and study how the function of the system can be restored after the disturbance. Further, the assessment can involve an evaluation of the level of vulnerability, and (if needed) an analysis of options for enhancing the robustness and/or resilience of the system. The assessment of an existing system involves checking its status or following up changes. A vulnerability assessment can, thus, facilitate the development of responses to possible crisis situations, and found the basis for prioritization between different alternatives to im-

prove system performance. The task of conducting an assessment can create an awareness of risk and vulnerability management in the organization and increases the motivation to work with these issues.

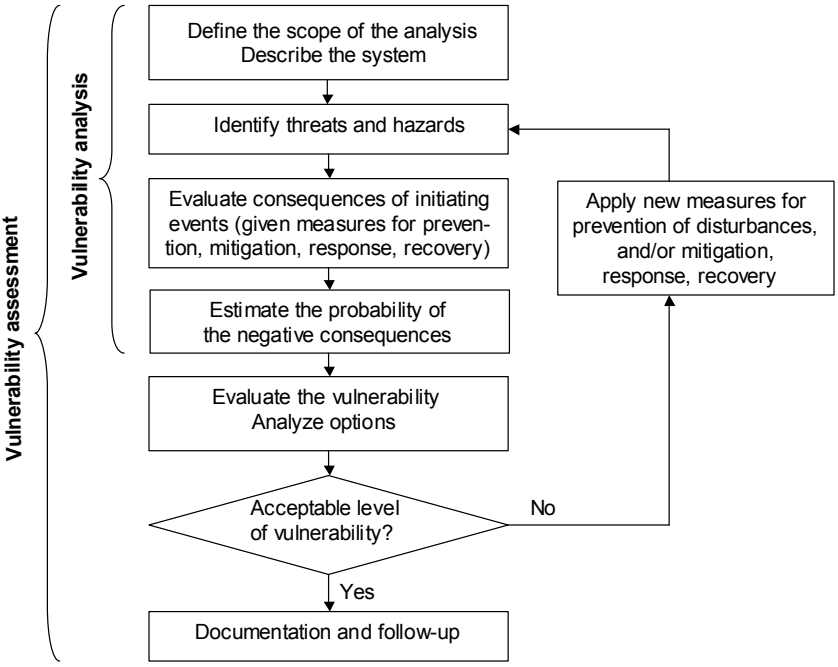


Fig 3.1 A framework for quantitative vulnerability assessment of infrastructures

3.4 Foundations of Vulnerability Analysis

A fundamental part of the vulnerability assessment is the vulnerability analysis, which can be captured by the following four questions (compare with Fig. 3.1):

1) What can go wrong?
2) What are the consequences?
3) How likely is it to happen?
4) How is a normal state restored?

The major difference between the risk and the vulnerability analysis is that the latter focuses on the whole *disturbance process* (the survivability of the system), and the major disturbances. For some initiating events (e.g. failure of technical components), it might be possible to estimate their fre-

quency. For other events, the conditional probability can be used (Eq. 2). Thus, the total probability is a sum where the terms consists of two parts: i) the probability that the initiating events A_i occur, and ii) the probability that this leads to consequences $Q > q$, i.e.

$$P(\max_{t\in T} Q(t) > q) = \sum_i P(A_i) \cdot P(\max_{t\in T} Q(t) > q \mid A_i). \tag{3}$$

A typical situation when analyzing the vulnerability of technical systems, especially when dealing with new technologies, is that there are few data of disturbances with severe consequences (a low probability high consequence, LPHC, problem). Useful information can be obtained from incidents (precursors), but it is seldom possible to use standard statistical techniques to estimate the vulnerability. Instead, mathematical models and/or experts' opinions have to be used. In summary, there are three principal ways to estimate the probability of occurrence of a negative event:

a) Statistical analysis of empirical disturbance (accident) data
b) Mathematical modeling combined with empirical component data
c) Expert judgments

Regarding the resulting negative *consequences* of an event, a similar division can be made. Within the engineering disciplines, analytical and numerical models play an important role in consequence analysis, among others for evaluating the consequences of fire, explosions, dispersion of chemical agents etc. There are advanced numerical models for static and dynamic analysis of power systems (e.g. optimal power flow). For example, Milano (2005) provides a description of an open source toolbox for design and analysis of small to medium size electric power systems.

Ordinary *statistical analysis* of *empirical* accident *data* is used extensively in studies of traffic and workplace accidents. The use of *mathematical modeling in combination with empirical component data* is well established in the nuclear and process industries (quantitative risk analysis or probabilistic safety analysis). *Expert judgments* are normally the primary sources of information in typical engineering risk analysis methods, and can be collected through more or less formalized methods (interviews, surveys, workshops etc.). Empirical data can also be combined with expert judgments with Bayesian statistical tools. Overall, the traditional *risk analysis offers a toolbox* of established quantitative, and semi-quantitative, methods for safety analysis of well-defined technical systems.

The rapid proliferation of *information and control systems* has increased the possibilities of optimizing, and controlling, industrial processes. Today, large technical systems are inherently so complicated that a layer of control, monitoring, and coordination is required for their normal opera-

tion. When software is combined with hardware to create programmable systems, the ability to assure conformity assessment through analysis, testing and certification becomes more difficult.

A fundamental problem in system studies lies in the fact that the response to all possible stimuli is not fully understood. Describing, and delimiting, a system as a first step in a vulnerability assessment is, thus, a daunting task. Uncertainties are experienced not only when it comes to the system itself, i.e. the *interactions* between the parts of the system, but also regarding the properties of the environment, i.e. the context. The interactions between different infrastructures, often referred to as *interdependencies*, are particularly important when dealing with critical infrastructure protection since infrastructures often act together to provide a service.

In the literature, critical infrastructures are typically portrayed as *complex systems*, but the meaning of the concept "complex" is often unclear. Commonly, the concept is used for *characterizing* the system, but it can also be a *metaphor* or analogy. The term complex can also be used to make an arbitrary distinction between something *perceived* as simple, and something perceived as complicated – the simple/complex dichotomy. Complexity, used as a metaphor, generally implies a critique against the traditional reductionist approaches and the predominant systems theory. Thus, it is a conception that synergies emerge when large sets of entities are brought together. Labeling a system complex, can also be a way of swiftly capturing properties considered to be the hallmarks of complexity, i.e. nonlinearity, adaptability, self-organization, emergence etc.

A variety of different measures would, hence, be required to capture all intuitive ideas about what is meant by complexity, and complexity, however defined, is not entirely an intrinsic property of the entity described; it also depends on who or what is doing the describing (Gell-Mann 1997). No attempts to make a formal definition of a complex system shall be undertaken, instead the author agree with Simon (1962):

> "Roughly, by a complex system I mean one made up of a large number of parts that interact in a nonsimple way. In such system, the whole is more than the sum of the parts, not in an ultimate, metaphysical sense, but in the important pragmatic sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole. In the face of complexity an in-principle reductionist may be at the same time a pragmatic holist."

Accordingly, the author argues that studies of critical infrastructures must rely on both detailed engineering modeling, and coarse modeling that focus on generic mechanisms. Existing methods for risk analysis can, to some extent, be adjusted and used in vulnerability analysis of infrastructure systems, but a major challenge is to further develop methods for analysis of complex systems (see examples in Sects. 3.5–3.7).

3.5 Example 1 – Statistical Vulnerability Analysis

Generation and trading of electricity in Sweden is carried out in a competitive environment, but Swedish grids are still regulated monopolies. The Swedish Energy Agency is responsible for ensuring that the grids are operated efficiently. As a part of the evaluation of tariffs, all utilities are obligated to report power outages to the Agency. Utilities typically publish compiled power outage data in annual reports, but seldom use statistical tools in the analysis. The aim of the author's study, presented in Holmgren and Molin (2005), is to explore the possibilities of using *statistical analyses of power outage data in vulnerability analysis* of electric power delivery (compare with approach a) in Sect. 3.4).

The vulnerability measure in Eq. (1) can be formulated as

$$P(Q > q) = 1 - F(q) = R(q)\,, \tag{4}$$

where $F(q)$ is the probability distribution function, and $R(q)$ is denoted the survivor function. For a continuous random variable, $F(q)$ is obtained by integrating the probability density function $f(q)$. The study includes data from the Swedish national transmission grid (153 observations from 11 years,), and the Stockholm distribution grid (Table 3.1). The power outage size Q is measured as the unserved energy (MWh), the power loss (MW), and the restoration time (h), i.e. there are six time series of power outage data.

Table 3.1 Power outage data from a Swedish distribution grid (1998–2003)

Cause	n	n_{max} [MWh]	n_{median} [MWh]	n_{Q3} [MWh]
Equipment failure[a]	325	3900	1.0	2.4
Unknown	55	106	0.6	1.4
Other[b]	45	9	1.6	2.9
Human factors[c]	41	11	0.3	1.3
Damage[d]	5	20	0.9	1.5
Nature/weather[e]	3	71	3.8	-
Lightning	2	65	33.1	-
All disturbances	476	3900	1.0	2.3

n number of recorded power outages, n_{max} largest power outage, n_{median} median power outage, n_{Q3} third quartile (75th percentile).

[a] Failure in technical equipment controlled by the utility.

[b] Technical, and human failures, outside the utility's responsibility.

[c] Failure by the utility's personnel.

[d] Deliberate attacks or sabotage.

[e] Natural hazards or adverse weather (except for lightning).

The *probability density functions f*(*q*) in all the data sets have *skewed shapes*, and the largest recorded power outage is 100 000 times larger than the smallest. This is a characteristic feature of time series of accident data from many areas, i.e. there are several minor accidents, but few major ones (the LPHC problem). Statistical distributions such as the log-logistic, and the lognormal, fit the data somewhat reasonable. Evaluations of probability plots show a tendency for the data to be heavier in the tails than in both these distributions (log-logistic cannot be rejected in hypothesis tests).

Recent studies of power outage data from the bulk electric systems in North America (data from the North American Reliability Council) show that the larger outages follow a *power law* (Chen et al. 2001; Carreras et al. 2000, 2004b). That is, there is good linear fit in a plot of the empirical cumulative survivor function $\ln(P(Q > q))$ versus the size of the power outages $\ln(q)$. The studies of the Swedish data also demonstrate that the power outage size follows a power law (see example in Fig. 3.2), where

$$P(Q > q) \sim A \cdot q^{-\beta} \quad (q \to \infty). \tag{5}$$

Since power law distributions have "heavy tails", the distribution allows for extremely rare events with extraordinarily large size (as compared to the standard normal distribution).

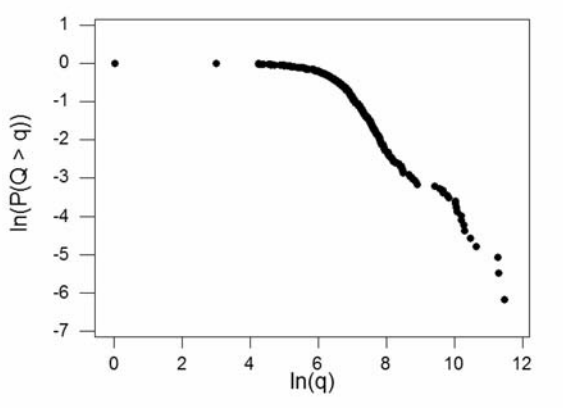


Fig. 3.2 Log-log plot of power outage data [power loss, MW] from the Stockholm distribution grid (1998–2003), i.e. $\ln(P(Q > q))$ versus for $\ln(q_n)$, $n = 1,\ldots,476$. The plot (and a regression analysis not displayed here) demonstrates that the distribution follows a power law for large q (Holmgren and Molin 2005).

Utilities can use information from *outage analysis* when deciding on equipment purchase or how to organize maintenance. Systems for reporting incidents and disturbances can give increased knowledge about how disturbances arise and how disturbances can be avoided. Further, statistical

analysis of outages data makes it possible to objectively follow-up the system performance, and to discover deficiencies that call for more detailed investigations. For both the studied Swedish power grids, there are no statistically significant shifts of the trend in the data – the outage size does not depend on the time. This can be an indication that the vulnerability of the systems not has changed considerably during the studied period.

3.6 Example 2 – Graph Theoretic Vulnerability Analysis

Complex systems can often in a useful way be described as networks, and networks can be represented as graphs. A *graph* $G = (V, E)$ can be defined as "a triple consisting of a vertex set $V(G)$, an edge set $E(G)$, and a relation that associates with each edge two vertices (not necessarily distinct) called its endpoints" (West 2001). Depending on what type of systems that is being observed, vertices and edges can be accentuated differently. In the following, the graphs will be undirected, and connected, which relates to the general structure (topology) of the network, whereas directed graphs relates to the actual flow of power in the network (given a specific operational scenario). Thus, the vertices can be generation units, stations, or users, and the edges can represent power lines.

Albert and Barabási (2002), and Dorogovtsev and Mendes (2002), review recent advances made in the field of graph theory and network analysis. A number of *statistical measures* have been proposed to characterize the structure of complex networks, and the following concepts are central:

- *Average path length:* the distance between two vertices is defined as the number of edges along the shortest path connecting them. In most complex networks there is, despite their often-large size, a relatively short average path length between any two vertices.
- *Clustering coefficient*: this measure captures the density of triangles in the graph. The clustering coefficient of a vertex is the ratio between the actual number of edges that exist between the vertex and its neighbors and the maximum number of possible edges between these neighbors.
- *Degree distribution* the number of edges connected to a vertex is called the degree. The degree distribution $P(k)$ of many empirical networks has a power law tail, $P(k) \sim k^{-\beta}$, where β is between 1 and 3 (Albert and Barabási 2002).

The studies of networks has given birth to several classes of abstract network models. Erdös and Rényi introduced the idea of *random graphs* in

the late 1950s. The simple random graph model combines low clustering with an exponential degree distribution. Watts and Strogatz introduced the so-called *Small World model* in 1998. This model combines high clustering and a short average path length (Watts and Strogatz, 1998). In 1999 Barabási and Albert presented the *Scale-free network model* that has a power-law degree distribution (Albert and Barabási 2002).

As far as the author knows, graph theoretic models have been used to study the following electric power grids (the same aspects have not been studied for all networks): the Western States transmission grid in the U.S. (Watts and Strogatz 1998; Amaral et al. 2000; Crucitti et al. 2004a), the North American grid (Albert et al. 2004), the Italian grid (Crucitti et al. 2004b; Rosato et al 2006), the French grid, the Spanish grid (Rosato et al 2006), and the Nordic transmission grid (Holmgren 2006).

Table 3.2 The structure of electric power transmission networks[a]

Network	C_{Actual}	C_{Random}	l_{Actual}	l_{Random}
The Western States power grid[b]	0.0801	0.00054	18.99	8.7
The Nordic power grid[c]	0.0166	0.00049	21.75	10.0

C_{Actual} Clustering coefficient (empirical network), C_{Random} Clustering coefficient (random graph of equivalent size), l_{Actual} Average path length (empirical network), l_{Random} Average path length (random graph of equivalent size).

[a] For formal definitions and algorithms, see Holmgren (2006).

[b] 4941 vertices and 6594 edges.

[c] 4789 vertices and 5571 edges.

In Holmgren (2006), an analysis of the *structural vulnerability* of the Nordic Interconnected grid and the Western States (U.S.) transmission grid is presented. Table 3.2 compares the structure of the power grids with random graphs of the equivalent size (calculations for the U.S. grid are also presented in Watts and Strogatz (1998)).

The Nordic grid is more scattered than the Western States (U.S.) grid, i.e. the average path length is larger and the clustering coefficient is lower. However, both transmission grids have a clustering coefficient significantly larger than the random graphs, and the average path length is more than twice as large as in the random graph. That is, the transmission grids show the "small world" phenomenon (the clustering coefficient is much larger than in the equivalent random graph, but the average path length is only somewhat larger in the power grids). Further, it is shown that both power grids have approximately exponential degree distributions, which also is a characteristic feature of the random graph. (A study of the degree distribution of the Western States grid was initially presented by Amaral et al. (2000).)

In the structural vulnerability analysis, failures are modeled by removing randomly chosen vertices of the graph (error tolerance). Attacks are realized through the removal of the vertices in decreasing degree order (attack tolerance). Two different attack strategies are studied: vertices are removed by their initial degree (number of connected edges), or the degree is recalculated after every removed vertex. The power grids are compared with two network models, i.e. a random graph and a scale-free network (see also Albert et al. (2000) and Holme et al. (2002)).

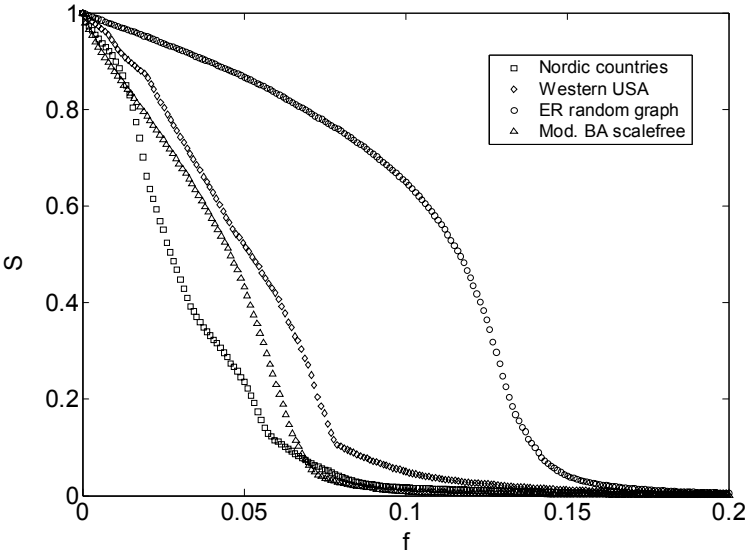


Fig. 3.3 Defragmentation of four different networks of approximately the same size. The vertices (fraction *f*) are removed in decreasing degree order (i.e. the vertex with most connected edges is removed first). After every removed vertex, the degree is recalculated. The relative size of the largest connected subgraph (component) *S* is used as a measure of the consequences of removing vertices, i.e. measure the attack tolerance of the network. The figure shows that the two electric power grids, and the scale-free network, are more sensitive to attacks than the random graph (Holmgren 2006).

Detailed data on the structure of the two transmission grids are restricted. Hence, it is not possible to separate vertices representing users from vertices representing other installations. Thus, different indirect measures are used to estimate the consequences of removing vertices in the network. The simulations confirm the results from previous studies above, and demonstrate that all studied networks disintegrate considerably faster when vertices are removed deliberately than randomly, i.e. the networks have a lower attack tolerance than failure tolerance. Further, the two electric power networks exhibit similar disintegration patterns, both for ran-

dom failures and attacks (Fig. 3.3). Also, it is shown that the scale-free network and the electric power grids are more sensitive to attacks than the random graph.

An important field of application for the vulnerability analysis is to evaluate alterations of an existing system. As an experiment, the graph of the Nordic power transmission grid is modified by incorporating two new edges (power lines) between Sweden and each country in the region, i.e. six new edges. The new power lines are positioned as in an internal study proposal from Svenska Kraftnät (SvK) – the utility that operates the Swedish national transmission grid. Comparing the augmented Nordic grid with the present Nordic grid, however, yields small, if no, visible changes in the error and attack tolerance (as analyzed here). Thus, a generic graph analysis, based on open-source data of the structure of the networks, is too simplistic for practical purposes.

Table 3.3 Examples of hazard and threat scenarios (Holmgren 2006)

Scenario	Description	Graph realization
Major technical failure	A major technical failure disables a station in the sub-transmission grid or the distribution grids.	Anyone of the vertices in the graph is removed with probability $p = 0.005$ (per year). Repair time: 12 h.
Snow-Storm	A snowstorm causes technical failures at the distribution level (overhead power lines).	Any two adjacent edges in the distributions grid are removed with $p = 0.01$. Repair time: 8 h.
Saboteur	This class of adversaries has a broad spectrum of motives, and can act irrational. The saboteur has little knowledge of the power system, does not have access to explosives, and is only capable of a single-entity attack.	Anyone of the vertices and edges in the distribution grids is a possible target. Attack by a rational (determined) saboteur: repair time 10 h. Attack by an irrational (opportunistic) saboteur: repair time 5 h.

In order to illustrate how the methodology can be applied in a more detailed evaluation of a system, a fictitious power delivery system is studied. A broad set of threat and hazard scenarios is represented as the removal of vertices and edges by introducing different repair times (a brief example is given in Table 3.3). The consequence of removing an entity in the graph might be a number of disconnected sink vertices (collectives of users). By assuming a load (MW) on each sink vertex, the consequence Q is measured as the *unserved energy* (MWh), here approximated as the power loss (MW) multiplied with the recovery time (h). The recovery time depends

on the repair times of the removed components. For all the scenarios, the measure $P(Q > q)$ is calculated (conditional vulnerability is used for attacks), and a relative comparison is made between three different tactics for upgrading the system: "Robustness" (strengthening the network by adding new edges), "Resilience" (shortening repair times), and "Combination" (a mix of the two other tactics).

As pointed out above, the study of abstract models can be a way of finding generic mechanisms, and increase the understanding of complex systems. Also, there are several practical reasons why studies of abstract network models of electric power systems can be a useful complement to the analysis of actual systems. Firstly, electric power grids, as other complex systems, are extremely large if modeled in detail, and the simulations will, therefore, be extremely demanding. Secondly, detailed data on electric power grids can seldom be obtained since they often are restricted. Thirdly, vulnerability assessment involves studies of antagonistic attacks. For security reasons, studies of attacks against authentic networks will most likely be classified.

However, the graph-based models described above are rather primitive, and a major drawback is that they do not capture how networks are operated. Electric power system analysis traditionally have a strong technical focus, including analysis of power flow, stability etc. for optimization of normal operations and emergency control (whereas the focus in this chapter is on "in extremis" states). For example, Salmeron et al. (2004) describe an analytical technique (an algorithm) to search for the worst-case disruptions in an electric power grid due to physical attacks. The terrorists' resources are specified as the number of people, and to interdict a power line, transformer station or sub station requires a given number of people.

Currently, there are no practically usable generic graph models of electric power grids. Holmgren and Thedéen (2006) use a simple analytical graph model to represent a distribution grid. The network is modeled as a random tree (branching process), and it is shown that failure in the network (removal of edges) results in a power outage size distribution that follows a power law (compare with Sect. 3.5). The branching process model captures the hierarchical nature of electric power grids, but at this stage it does not include clustering (the clustering coefficient $C = 0$ in a tree since there are no cycles).

Major power outages typically include *cascading failures* in electric power transmission grids, i.e. multiple failures that are the direct result of a common or shared root cause (UCTE 2003; U.S.-Canada Task Force 2004). Given a lightly loaded power system, there is a very low likelihood that a trip in a power line will cause a power outage. As the load increases, more dependent failures occur, and at some critical load, a trip in a power

line might cause an instability that cascades in the network, and eventually resulting in a major blackout.

There are several different approaches to studying cascading failures in power systems, see Dobson et al. (2005) for an overview of this subject. For example, Carreras et al. (2004a), use a DC load flow approximation, and standard linear programming optimization, to represent cascading transmission line overloads. Motter and Lai (2002) as well as Crucitti et al. (2004b) use graph models (simulation) that do not consider the flow in networks. There are also analytical models to study cascading failures, and Dobson et al. (2004) presents a branching process model for approximating the propagation of failures in a transmission grid.

In summary, the author believes that it is vital to improve the understanding of the relationship between *dynamics* and vulnerability of complex networks. Thus, vulnerability analysis of electric power networks would benefit greatly from more cross-fertilization between electric power engineering, and the network modeling and simulation of complex systems as introduced here.

3.8 Example 3 – Game Theoretic Vulnerability Analysis

Antagonistic attacks are typically analyzed using conditional probabilities (Eq. 2). To use the probability concept when dealing with planned attacks is, however, problematic. The measures applied to protect the infrastructure will affect the antagonist's course of action (assuming an informed adversary). Changes in how the defender perceives that the opponent will act, will again affect how the defense is allocated, which once more can affect the antagonist's behavior etc. There is an *interaction* between the attacker and the defender. Therefore, studies of attacks embrace a game situation rather than a decision situation. In defense analysis, *game theory* is widely used to analyze the effects of selecting alternative strategies to achieve a military objective (Shubik and Weber 1981). Games are used for planning, education, and for generating knowledge. Penetration testing ("red teaming") is conducted to seek out technical and structural weaknesses in computer systems, and for studying attack approaches and consequences of attacks.

Paté-Cornell and Guikema (2002) presents a model based on probabilistic risk analysis, and elements of game theory, for setting priorities among threats and among countermeasures. Bell (2003) studies the vulnerability of networks, and a game is set up between a router, who seeks to minimize the travel cost for data packets (or vehicles) by choosing routes in the net-

work, and an antagonist, who seeks to maximize the travel cost by destroying edges. Bier et al. (2005) apply elements of game theory and network reliability analysis to identify optimal strategies for allocating resources to defend idealized systems against attacks.

In Holmgren et al. (2006), the interaction between an attacker and a defender of a power system is modeled as a game. In a numerical example (using a maximum-flow lossless network model for calculating the consequences of attacks), the performance of different defense strategies against a number of attack scenarios is studied. An attack results in disabled elements in the network, which in turn may lead to loss of power for users (sink vertices). The total consequence of an attack is measured as the energy loss (MWh), which is approximated as the power loss multiplied with the recovery time.

In the model, the defender can only spend resources on increasing the component protection (e.g. fortification), and/or decreasing the recovery time after an attack (e.g. repair teams), i.e. the defense budget $c_{\text{total}} = c_{\text{prevent}} + c_{\text{recovery}}$. Every element i (vertices and edges) in the network has a protection described by the parameter p_i. This parameter corresponds to the probability that an attack against element i fails. The protection p_i of element i is a function of the resources c_i spent on protecting that element. The defender distributes the resources for protection between the N elements in the network. The repair time of element i depends on the resources spent on recovery, as well as the type of the disabled element, and the attack method. In the model, it is assumed that the defender has a basic recovery capacity for maintenance and for repairing minor failures. Thus, the relative contribution of spending extra resources on recovery is studied. In summary, the total allocation of defense resources is described by the vector $\mathbf{c} = (c_1,\ldots, c_N, c_{\text{recovery}})$.

The attack model only considers qualified antagonists. That is, determined, well-informed, and competent antagonists with access to enough resources to perform a successful attack against an electric power system. The antagonist is allowed to randomize between which targets to attack, and r_j correspond to the probability that target j is attacked (a target can consist of more than one element in the network), given that an attack is made. The vector $\mathbf{r}$ of dimension M then describes the mixed strategy, and three different classes of attack strategies are considered:

- *Worst-Case Attack:* The antagonist chooses the target that maximizes the expected negative consequences of the attack.
- *Probability-Based Attack:* The antagonist tries to maximize the probability that the outcome of an attack is over a certain magnitude q, i.e. $P(Q > q)$.

- *Random Attack:* The antagonist chooses the attack target randomly, and each target is attacked with equal probability.

An attack scenario is constructed by specifying the class of attack strategy, and a few additional parameters that captures tactics and modes of operation. The aim is to make the attack scenario more realistic by adding a few conditions and restrictions (e.g. regarding the amount of damage that can be inflicted to the targeted elements)

The interaction between the defender and the antagonist is described as a two-player zero-sum game, where, simultaneously, the defender chooses an allocation of defense resources, and the antagonist chooses a target to attack. Consequently, it is assumed that the defender's payoff is the negative value of the attacker's payoff. The situation where the attacker tries to maximize and the defender tries to minimize the total expected damage can, thus, be translated into an optimization problem. The game theory model has deliberately been kept simple, and it is assumed that both players have perfect information about the system, and the resources and preferences of the other.

In a simple numerical example (using a stylized version of the national Swedish transmission network) the performance of different defense strategies against a number of attack scenarios is studied. For this example, it is possible to find an optimal allocation between protection and recovery for the given scenarios (Fig. 3.4). This allocation depends on the total amount of resources c_{total} and the attack scenario. During an extreme situation there are more elements whose failure will cause large negative consequences compared to the normal situation. As a result, in this situation it is more effective to spend a larger proportion of the resources on recovery than during the normal situation.

It is not possible to find a *dominant defense strategy* in the numerical example. That is, a defense strategy with lower expected negative consequence than every other defense strategy against every attack scenario. A defense optimized against the Worst-Case Attack strategy will not necessarily provide an optimal defense against other attack scenarios (e.g. a scenario involving a Probability-Based Attack strategy). It is possible to use a number of statistical methods to give a ranking of the different defense strategies, and a few different ways of comparing the different defense strategies against each other are discussed in the paper.

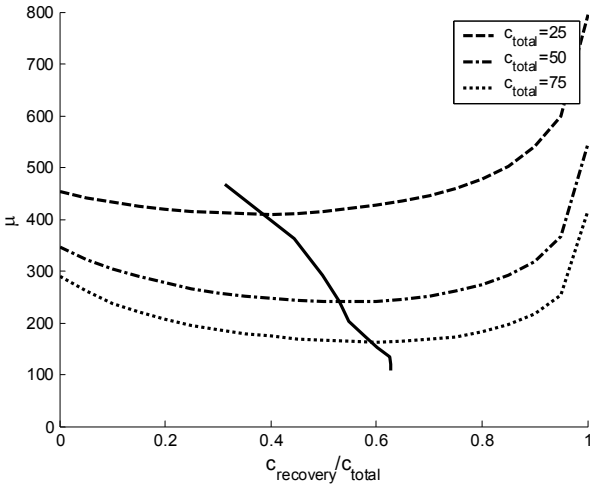


Fig. 3.4 Numerical example from a game theoretic vulnerability analysis model presented in Holmgren et al. (2006). The figure shows the balance between resources for protection and recovery for a given pair of defense and attack scenarios. The dotted lines display the expected negative consequences μ for three different total amount of resources c_{total} as a function of the fraction $c_{\text{recovery}}/c_{\text{total}}$. The solid line shows the optimal distribution between protection and recovery for different budgets c_{total}, i.e. the minimum of the dotted lines. Extra calculations have been made to find the optimal distribution for c_{total} between the horizontal lines.

In conclusion, it is well known that theoretical results in game theory depend significantly on how the game situation is modeled (the set of players, the set of strategies for each player, the choices that each player can make, the set of payoffs corresponding to the utility each player can receive etc.). Modeling antagonistic attacks against infrastructures, the information is very limited, and it becomes difficult to exactly specify the structure of the game. However, the author believes that using concepts and general models from game theory is a very powerful way of framing the problem.

3.8 Vulnerability Evaluation

3.8.1 Vulnerability Evaluation Criteria and Strategic Options

The *vulnerability evaluation* (compare with Fig. 3.1) can be based on different decision criteria:

- *Technology* based criteria (e.g. best practice or best available technology)
- *Right* based criteria (e.g. formulated in prescriptive standards or regulations given as quantitative limits)
- *Utility* based criteria (e.g. cost-benefit/cost-effectiveness analysis).
- *Combination* criteria

If the level of vulnerability cannot be accepted, there are several strategic options. It might be possible to avoid, or prohibit, certain activities (*avoidance*), or a choice, intentional or unintentional, can be made not to take any actions at all (*retention*). That is, to bear potential negative consequences within the normal activities. Further, actions or measures improving the protection of the infrastructure can be employed (*reduction*). The responsibility can also be transferred to another entity (*distribution*), e.g. via insurance, or a combination of retention and transfer (*sharing*) can be used, e.g. forming joint ventures.

An infrastructure operator might face threats with a potential of causing extremely large negative societal consequences. In a commercial contract, events such as these can be covered by a *Force Majeur* clause (if the event, and its effects, is considered to be outside the operator's possibility to control, the operator might be relieved of further responsibility). To ensure the survival of the company, and to hedge against commercial loss, a private infrastructure operator might use some insurance solution. However, to fill the gap between national security and risk management in private organizations, some form of *public commitment* is often required.

3.8.2 Options for Electric Power Systems Protection

Crisis management consists of a number of phases, for example: prevent, mitigate, response, recover, and learn. Measures for the *prevention* of failures and attacks aims at reducing the likelihood, or avoiding, that an event occurs. *Mitigation* aims at minimizing the negative consequences of an event. *Response* includes measures performed during the acute crisis phase in order to minimize the negative consequences of an event. Finally, *recovery* involves all measures carried out to bring back the system to a normal state after an event.

A general principle can be to first try to prevent a systems from degenerating into alert and emergency states, but if this does occur, it is important to minimize the disturbance, and restore normal conditions as quickly as possible. However, to prevent major power disturbances is generally considered to be complicated, and requires substantial resources. For ex-

ample, the Swedish transmission grid consists of some 15 000 km of overhead power lines, localized mainly in rural and uninhabited areas. Thus, it is *not* economically, or technically, achievable to fully eliminate the vulnerability of the Swedish power system in relation to antagonistic attacks. Consequently, for some threats, the solution can be to allocate more resources to response and recovery (compare with Sect. 3.7).

Prevention and Mitigation

Some general tactics for *prevention* and *mitigation* are: barriers (to confine/restrict a condition with potential for harm); redundancy (to improve system availability through additional, identical, components); diversity (applied to equipment, functions, and staff); training, quality control, and procedures review; preventive maintenance; monitoring, surveillance, testing and inspection (Parry 1991).

Electric power transmission grids are commonly designed and operated according to the deterministic "*N* - 1 Criterion". That is, the whole system must be capable of operating normally even when a major failure occurs. Measures to avoid failures in technical systems have traditionally been concerned with the safety perspective, but the tactics listed above are also suitable for creating physical security. Also, there is a variety of security mechanism that is designed to detect, prevent, or recover from a cyber attack, e.g. firewalls, Intrusion Detection Systems, and anti virus software.

Response and Recovery

The response to a power outage can be based on the same principles as normal electric power system operations. The *emergency control* involves automatic countermeasures to cope with instabilities in the power grid (e.g. load shedding can be implemented to manage loss of power generation), and the use of system monitoring tools (computer based early-warning systems) to keep the system from degenerating further.

Power systems *restoration* includes determining the detailed state of the system, preparing the equipment for restoration to service, reintegrating and rebuilding the system, and balancing generation and load as they, in a controlled manner, are brought back to their normal level. A general tactical choice is between the "build-down" approach (i.e. reenergizing the bulk power network before resynchronizing most generators), and the "build-up" approach (i.e. restoring islands that will then be mutually interconnected). The "build up" approach is more common and usually selected in a scenario involving a complete system collapse (Ancona 1995; Adibi and Fink 1994).

3.9 Concluding Remarks

The crisis management of large-scale power outages demands coordinated actions between countries, and is therefore of interest to the international community. The process will involve stakeholders both from public and private organizations. Even though the transnational terrorism and the cyber threats are alarming, major blackouts in recent years show that adverse weather and technical failures need consideration.

Critical infrastructure protection demands a holistic view; both technical and non-technical factors are of great importance. Thus, a vulnerability assessment methodology based on *multiple perspectives* is recommended. *Proactive work* is needed in order to assure that the infrastructure systems will be able to supply the services that a modern society relies on. A general principle can be first to try to prevent the systems from degenerating into alert and emergency states, but if this does occur, it is important to minimize the extent of the disturbance, and restore normal conditions as quickly as possible.

The preferred vulnerability analysis approach depends on the *objective of the analysis*, but also on the *available information* about the system. The traditional risk analysis offers a toolbox of well-established quantitative methods, and can to some extent be used to analyze the vulnerability of the technical systems that form the infrastructure. However, recent advances in network modeling and simulation, and also game theoretical approaches, should be taken into account.

Even if a systematic vulnerability assessment is conducted, decisions on critical infrastructure protection will involve a great deal of *uncertainty*. Commonly proposed solutions are to take decisions successively (i.e. using *adaptive strategies*), and to develop the *ability to act on unexpected situations* as they emerge (e.g. through the use of games as a learning and planning tool). Other recommendations are that uncertainties relevant to decision situations should be made explicit and understandable to the decision makers, and that a vulnerability assessment should include some form of *sensitivity analysis*.

Acknowledgements

The author would like to thank the following persons for valuable discussions and comments: T. Thedéen, S. Molin, L.-G. Mattsson, S. Arnborg, H. Christiansson, E. Jenelius, and J. Westin.

Financial support from the Swedish Emergency Management Agency (contract no. KBM 0054/2002) is gratefully acknowledged.

References

Adibi MM, Fink LH (1994) Power system restoration planning. IEEE Transactions on Power Systems 9: 22–28

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Reviews of Modern Physics 74: 47–97

Albert R, Albert I, Nakarado GL (2004) Structural vulnerability of the North American power grid. Physical Review E 69: 025103(R)

Albert R, Jeong H, Barabási A-L (2000) Error and attack tolerance of complex networks. Nature 406: 378–381

Amaral LAN, Scala A, Barthélemy M, Stanley HE (2000) Classes of small-world networks. Proceedings of the National Academy of Sciences 97: 11149–11152

Ancona JJ (1995) A framework for power system restoration following a major power failure. IEEE Transactions on Power Systems 10: 1480–1485

Bell MGH (2003) The use of game theory to measure the vulnerability of stochastic networks. IEEE Transactions on Reliability 52: 63-68

Bier WM, Nagaraj A, Abhichandani V (2005) Protection of simple series and parallel systems with components of different values. Reliability Engineering & System Safety 87: 315-323

Carreras BA, Newman, DE, Dobson I, Poole AB (2000) Initial evidence for self-organized criticality in electric power system blackouts. Proceedings of the 33rd Hawaii International Conference on System Sciences, Hawaii

Carreras BA, Lynch VE, Dobson I, Newman DE (2004a) Complex dynamics of blackouts in power transmission systems. Chaos 14: 643-652

Carreras BA, Newman DE, Dobson I, Poole AB (2004b) Evidence for self-organized criticality in a time series of electric power system blackouts. IEEE Transactions on Circuits and Systems--I: Regular papers 51: 1733-1740

Chen J, Thorp J, Parashar M (2001) Analysis of electric power system disturbance data. Proceedings of the 34th Hawaii International Conference on System Sciences, Hawaii

Crucitti P, Latora V, Marchiori M (2004a) Model for cascading failures in complex networks. Physical Review E 69: 045104(R)

Crucitti P, Latora V, Marchiori M (2004b) A topological analysis of the Italian electric power grid. Physica A 338: 92–97

Dobson I, Carreras BA, Newman DE (2005) A loading-dependent model of probabilistic cascading failure. Probability in the Engineering and Information Sciences 19: 15-32

Dobson I, Carreras BA, Newman DE (2004) A branching process approximation to cascading load-dependent system failure. Proceedings of the 37th Hawaii International Conference on System Sciences, Hawaii

DoE (2002a) Vulnerability assessment methodology: electric power infrastructure. U.S. Department of Energy (DOE), Washington DC

DoE (2002b) Energy infrastructure risk management checklists for small and medium sized facilities. U.S. Department of Energy (DOE), Washington DC

Doorman GL, Uhlen K, Kjølle GH, Huse ES (2006) Vulnerability analysis of the Nordic Power System. IEEE Transactions on Power Systems 21: 402-410

Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Advances in Physics 51: 1079–1187

Einarsson S, Rausand M (1998) An approach to vulnerability analysis of complex industrial systems. Risk Analysis 18: 535–546

Gell-Mann M (1997) The simple and the complex. In: Alberts D, Czerwinski T (eds) Complexity, global politics, and national security. National Defense University, Washington, DC

Hansson SO, Helgesson G (2003) What is stability? Synthese 136: 219–235

Holme P, Kim BJ, Yoon CN, Han SK (2002) Attack vulnerability of complex networks. Physical Review E 65: 056109

Holmgren ÅJ (2006) Using graph models to analyze the vulnerability of electric power networks. Risk Analysis 26: 955–969

Holmgren ÅJ, Molin S (2005) Using disturbance data to assess vulnerability of electric power delivery systems. To appear (accepted October 2005) in Journal of Infrastructure Systems

Holmgren ÅJ, Thedéen T (2006) Structural vulnerability analysis of electric power networks. Submitted manuscript

Holmgren ÅJ, Jenelius E, Westin J (2006) Optimal defense of electric power networks against antagonistic attacks. To appear (accepted October 2006) in IEEE Transactions on Power Systems

Høyland A, Rausand M (1994) System reliability theory: models and statistical methods. Wiley, New York

IEC (1995) Dependability management – part 3: application guide – section 9: risk analysis of technological systems. International Electrotechnical Commission (IEC), Geneva

Jenelius E, Petersen T, Mattsson L-G (2006) Importance and exposure in road network vulnerability analysis. Transportation Research Part A 40: 537–560

McEntire DA (2005) Why vulnerability matters: exploring the merit of an inclusive disaster reduction concept. Disaster Prevention and Management. 14: 206-222

Milano F (2005) An open source power system analysis toolbox. IEEE Transactions on Power Systems 20: 1199-1206

Motter AE, Lai Y-C (2002) Cascade-based attacks on complex networks. Physical Review E 66: 065102

Paté-Cornell E, Guikema S (2002) Probabilistic modeling of terrorist threats: a systems analysis approach to setting priorities among countermeasures. Military Operations Research 7: 5-20

Parry GW (1991) Common cause failure analysis: a critique and some suggestions. Reliability Engineering & Systems Safety 34: 309–326

Rosato V, Bologna, S, Tiriticco F (2006) Topological properties of high-voltage electrical transmission networks. Electric Power Systems Research (in Press)
Salmeron J, Wood K, Baldick R (2004) Analysis of electric grid security under terrorist threat. IEEE Transactions on Power Systems 19: 905-912
Shubik M, Weber RJ (1981) Systems defense games: Colonel Blotto, Command and Control, Naval Research Logistics Quarterly 28: 281-287
Simon HA (1962) The architecture of complexity. Proceedings of the American Philosophical Society 106: 467–482
UCTE (2003) Interim report of the investigation committee on the 28 September blackout in Italy. The Union for the Co-ordination of Transmission of Electricity (UCTE), Brussels
U.S.-Canada Task Force (2004) Final report on the August 14th 2003 blackout in the United States and Canada: causes and recommendations. U.S.-Canada Power System Outage Task Force
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442
West DB (2001) Introduction to graph theory. Prentice Hall, Upper Saddle River
White House (2002) The national strategy for homeland security. Washington DC

4 Spatio-Temporal Models for Network Economic Loss Analysis Under Unscheduled Events: A Conceptual Design

Jong Sung Lee[1] and Tschangho John Kim[2]

[1] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, USA; Email: jonglee@ncsa.uiuc.edu
[2] Department of Urban and Regional Planning, University of Illinois at Urbana-Champaign, USA; Email: tjohnkim@uiuc.edu

4.1 Introduction

The damages and losses caused by unscheduled events such as earthquakes, floods, and other major natural disasters have sudden and significant impacts on the economies of regions where these events occur. The impacts of damage on production facilities and lifelines (e.g. utility lines and transportation networks) may spread across several regions via import-export relationships and have serious economic impacts on even distant regions far from the location of the event.

Economic impacts from unscheduled events stem not only from damages and direct losses, but also from recovery and reconstruction activities. To recover and reconstruct facilities and lifelines damaged by unexpected events using investment or government financial aid, both the direct and indirect economic impacts from the events need to be measured in regional and interregional contexts. Direct economic impacts are defined as direct changes in production and demand due to the disruption of production facilities and lifelines from unexpected events; indirect economic impact is defined as the change in other sectors due to the change in a sector based on inter-industry relationships (Kim, Ham, and Boyce 2002; Ham, Kim, and Boyce 2005a, 2005b; Sohn et al. 2002; Sohn et al. 2003).

Issues that are not typically studied, however, include the temporal configuration in network economic loss analysis. Previous research analyzed the network loss for a certain given year (Kim, Ham, and Boyce 2002;

Ham, Kim, and Boyce 2005a, 2005b; Sohn et al. 2002; Sohn et al. 2003). The results, however, do not reflect the spatio-temporal changes even though the impact of unscheduled events, such as earthquakes, changes over time. In this chapter, the temporal configuration of data is modified from yearly to quarterly. Annual final demand and total output becomes quarterly. Changes in temporal configuration, however, raise two issues: 1) how to manage the spatio-temporal data, and 2) how to improve the static equilibrium model to reflect the spatio-temporal changes.

This chapter suggests solutions to these problems. The first problem can be resolved through designing and implementing a spatio-temporal data model. Towards developing a spatio-temporal model, existing spatio-temporal models are reviewed in Section 4.2. Based on the review, the data model for network loss analysis is proposed in Section 4.3. The methodology for modifying the static equilibrium model to reflect the spatio-temporal changes is suggested in Section 4.4. The data and preliminary analysis results are discussed in Section 4.5. Finally, future directions for developing a physical design for implementation are discussed in Section 4.6.

4.2 Review of Spatio-Temporal Data Models

Since Hägerstrand (1970) introduced the notion of time as a third dimension in spatial analysis, various researchers have included the dimension of time in their analyses (Peuquet 2001, El-Geresy et al. 2002). Spatio-temporal data models allow us to have a better understanding of the spatio-temporal changes in a dynamic environment. Spatio-temporal data models enable us to trace and analyze historical changes, and to forecast and estimate future changes (El-Geresy et al. 2002).

El-Geresy et al. (2002) propose three organizational bases for classifying spatio-temporal data models: ‘Time’, ‘Space’, and ‘Feature’. They are expressed as axes that construct the ‘data space’ as shown in Figure 4.1. Based on this classification, spatio-temporal data models can be classified into three categories such as ‘Space based approach’, ‘Time based approach’, and ‘Feature based approach’ depending on which organizational basis is facilitated. We propose a fourth category, ‘Multiple bases approach’ for the models that have multiple organizational bases. Table 4.1 shows the taxonomy of the spatio-temporal data models.

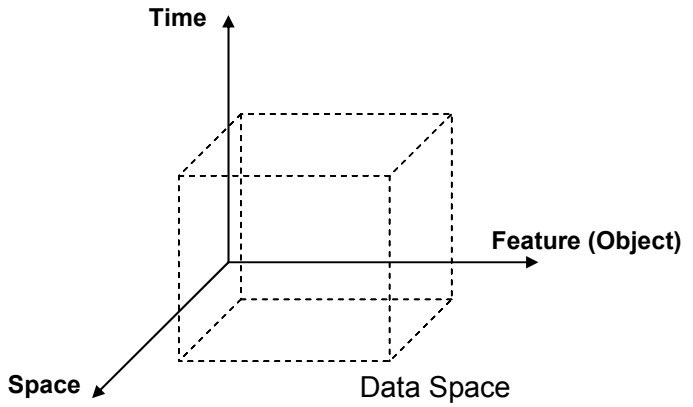


Fig. 4.1. Problem space with three organizational bases (El-Geresy et al. 2002, p. 192)

Table 4.1. Taxonomy of spatio-temporal data models

Approaches	Models	Articles
Space based approach	Location based model	(Langran 1988;1992)
Time based approach	Snapshot model	(Langran 1992) (Armstrong 1988) (Yuan 1994)
	Event based model	(Peuquet and Duan 1995)
	Process based model	(Claramunt and Thériault 1995) (Pang and Shi 2002)
Feature based approach	Amendment vector model	(Langran 1989; 1992)
	Space-time composite model	(Langran 1992)
	Spatio-temporal object model	(Worboys 1994a; 1994b)
Multiple Bases approach	Triad model	(Peuquet 1994)
	Three domains model	(Yuan 1994)

Abraham and Roddick (1999) and Peuquet (2001) proposed the object-oriented modeling approach for development of spatio-temporal models. The object-oriented modeling approach itself is not a model but a model design method. Therefore, even if the models use the same object-oriented modeling approach, the characteristics of the models might be different.

All models are extensions of traditional representation data models, such as raster and vector models, in terms of their representational method. Most are extensions of existing DataBase Management Systems (Peuquet 2001).

Models that are organized according to Space (or Location) are a typical example of a raster model. Because of the model's simplicity, implementing, querying and accessing the model is simple and efficient.

As an extension of vector based models, those models that have Feature as the organizational basis have been developed. The feature is defined as a simple feature (line, point, polygon, etc.) (Langran 1989). The space-time composite model extends the simple feature to the feature that represents geographical change in two-dimensional space (Langran 1992). The space-time composite, finally, is extended to the spatio-temporal object that represents spatio-temporal change by extending the two-dimensional space to three-dimensional space with the time dimension (Worboys 1994a, 1994b).

The advantage to this approach is that the topological relationship and integrity of individual features are explicitly maintained (Peuquet and Duan 1995). However, the redundancy of features could be an issue depending on the concept of features (Yuan 1994).

Unlike the Location based and the Feature based models, models with Time as the organizational basis are not confined to raster models and can be adopted by both raster and vector based models. However, the spatio-temporal changes involve the issue of maintaining the integrity of topological relationships. The model by Claramunt and Thériault (1995) deals with this issue by employing the versioning technique. Pang and Shi (2002) handle the issue by using the Voronoi model.

The notable point of departure from other Time based models is the concept of events. The concept of events has evolved from a moment when the event occurs (Peuquet and Duan 1995) to a set of processes of the event itself (Claramunt and Thériault 1995; Pang and Shi 2002). Time based models with events provide an efficient method for spatio-temporal queries (Peuquet and Duan 1995).

The last category consists of the models that have multiple organizational bases. For the Triad model (Peuquet 1994), the bases are Space (where), Time (when), and Feature (what). Similarly, the three domains model (Yuan 1994) has temporal, spatial, and semantic (aspatial) domains as organizational bases. The models can incorporate both raster and vector representation models. Also, the models conceptually allow queries according to both location and feature. In the following section, spatio-temporal data models for network loss analysis are suggested.

4.3 Conceptual Design for Spatio-Temporal Data Model

4.3.1 Characteristics of Data in Network Loss Analysis

In this section, the characteristics of data used in network loss analysis in terms of space and time are investigated in order to design a spatio-temporal data model for network loss analysis. The characteristics are classified into three categories: spatial characteristics, temporal characteristics, and spatio-temporal characteristics.

Spatial characteristics:

Data for network loss analysis consists of three simple spatial features such as points, lines, and polygons based on the Vector-based spatial model. These features support transportation analysis and multi-regional input output analysis.

Data that supports transportation analysis contains highway and railway networks. Each network consists of lines (called links) and points (called nodes). Among the nodes, there are centroids which represent earthquake analysis zones (EQAZs). Secondly, the 83 EQAZs, which are polygons, support the data for multi-regional input output analysis. In other words, the data of multi-regional input output analysis can be effectively visualized via EQAZs. The centroids provide the linkage between transportation analysis and multi-regional input output analysis. In addition, the epicenters, which allow an analysis of the earthquake impacts, are represented by point features.

In short, the important spatial characteristic of network loss analysis is that it is a vector-based (or feature based) model consisting of three simple features such as points, lines, and polygons.

Temporal characteristics:

The data for network loss analysis has two kinds of temporal elements: the data for transportation and multi-regional input output analysis have three months as a unit and the historical earthquake data has an irregular time interval.

Spatio-temporal characteristics:

Spatio-temporal characteristics are related to the spatio-temporal changes in network loss analysis in this chapter. The commodity flow on the network is to be updated (or changed) every 3 months. Multi-regional input output analysis data, such as total output, is to be changed quarterly

as well. Note that the commodity flow on the network is changed since the multi-regional input output analysis data is changed over time.

In network loss analysis, the major analysis is calculating the economic impact (or loss) caused by disruptions on the transportation network due to the earthquake. Therefore, disruptions or damages of the links change over time based on the recovery scheme (or function).

In addition, the spatial configuration is not changed over time while attributes such as commodity flows and total output are changed quarterly; this is a unique spatio-temporal characteristic.

4.3.2 Suitability of Spatio-Temporal Data Models

Considering the characteristics examined above, the feature based approach is judged to be suitable for network loss analysis due to the following reasons. Since network loss analysis consists of various feature types such as points, lines and polygons, the feature should be the basis for organizing the spatio-temporal data or changes. In addition, the data is in a vector-based format. Thus, the space based approach and multiple bases approach are judged to be unsuitable for network loss analysis.

Next, the changes of attributes of features are of major concern in this chapter. Since the frequency of changes is not often and is a fixed interval, managing the time components is a relatively minor issue. In other words, compared to traffic data or hydrological data, the spatio-temporal changes of network loss analysis occur in long-term intervals (3 months). Moreover, the time interval of the changes is fixed. This makes the time based approach unsuitable.

Table 4.2. Suitability of spatio-temporal data models

Approaches	Models	Suitable?
Space based approach	Location based model	No
Time based approach	Snapshot model	No
	Event based model	No
	Process based model	No
Feature based approach	Amendment vector model	Maybe
	Space-time composite model	Yes
	Spatio-temporal object model	Yes
Multiple Bases approach	Triad model	No
	Three domains model	No

Among the models in the feature based approach, the space-time composite (Langran 1992) and spatio-temporal object models (Worboys 1994a, 1994b) are adapted to the spatio-temporal data model for network loss

analysis. Since the Space-Time composite model is an improved version of the Amendment vector model, the amendment vector model is disregarded.

4.3.3 Spatio-temporal Data Models for Network Loss Analysis

Since the network loss analysis data are of two different kinds in terms of spatio-temporal characteristics: 1) features that have attribute changes over regular time interval, 2) point-features that have attribute changes and spatial changes over irregular time interval, two models are utilized. The data of network loss analysis can be classified into these two categories as shown in Table 4.3.

For the data in category 1, the Space-Time composite model is applied since it can handle the attribute changes over time efficiently with minimal data redundancy (Langran 1992). Similarly, Shaw and Xin (2003) adapt this model to their spatio-temporal data model for the exploratory analysis of interaction between transportation and land use.

The spatio-temporal object data model (Worboys 1994a; 1994b) is facilitated for the data in category 2. Since the epicenter has simple attributes such as location, moment of magnitude and time of occurrence, and most epicenters have different locations, it does not raise the data redundancy problem of the spatio-temporal object model, which Yuan (1994) has mentioned.

Table 4.3. Two types of data in terms of spatio-temporal characteristics

	Category 1	**Category 2**
Attribute changes over time	Yes	Yes
Spatial changes over time	No	Yes
Time interval	Regular	Irregular
Data	Network (points and lines) EQAZ (polygon)	Epicenters (points)

Those models are designed via an object-oriented modeling technique. The unified modeling language (UML) diagram of the spatio-temporal model of network loss analysis is shown in Figure 4.2. This model contains three components: Space-Time Composite (STC) components, Snap Shot (SS) components, and Spatio-Temporal Object (STO) components. Note that the name of each class begins with the abbreviation (e.g. STO_Epicenter).

According to the methodology developed by Shaw and Xin (2003), Space-Time Composite components are built from the Snap Shot components. Each snap shot layer of data contains the data at a certain time in-

terval (e.g. transportation data at first quarter of the year 2000). It becomes a Space-Time Composite in the STC components.

In the STO components, an object of STO Epicenter class is considered as a spatio-temporal atom (Worboys 1994a; 1994b). The Epicenters class is the collection of these spatio-temporal atoms.

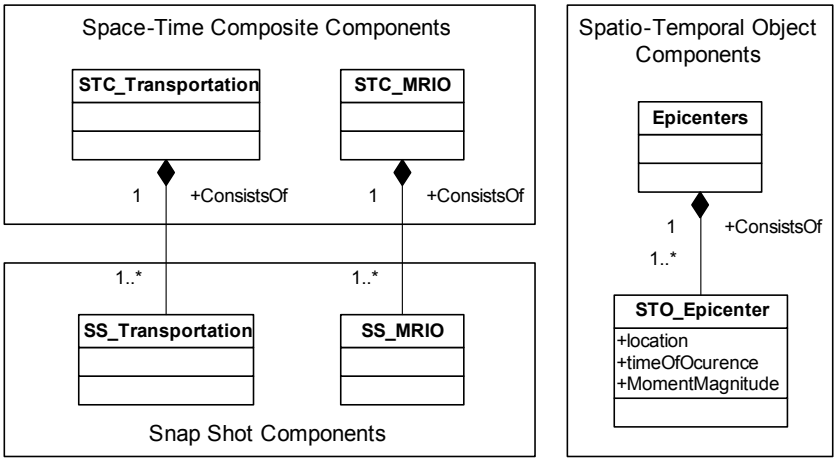


Fig. 4.2. UML diagram for spatio-temporal data model

Table 4.4. Related ISO standards

	Number	Title
Spatial feature	ISO 19107:2003	Geographic information – Spatial schema
	ISO/DIS 19125-1	Geographic information - Simple feature access - Part 1: Common architecture
Time	ISO 19108:2002	Geographic information – Temporal schema
Network	ISO/DIS 19133	Geographic information - Location based services tracking and navigation
	ISO/DIS 19134	Geographic information - Multimodal location based services for routing and navigation

The spatio-temporal data models can be enriched by adopting ISO standards: ISO 19107:2003 and ISO/DIS 19125-1 for spatial feature modeling, ISO 19108:2002 for temporal object modeling, and ISO/DIS 19133 and ISO/DIS 19134 for network feature modeling (see Table 4.4) since those standards provides the conceptual framework of the data models and those are agreed and developed by internationally well-known experts.

4.4 Spatio-Temporal Analysis Models of Network Economic Loss

In order to capture the dynamics of nationwide economic impacts of infrastructure on the transportation network due to unscheduled events, the Spatio-Temporal Analysis Model (STAM) is developed based on the analysis models developed by Kim, Ham, and Boyce (2002); Ham, Kim, and Boyce (2002); Ham, Kim, and Boyce (2004); Sohn et al. (2002) and Sohn et al. (2003). The analysis model consists of the Integrated Commodity Flow Model (ICFM), Final Demand Loss Model (FDLM), and Multiregional Input Output (MRIO) model.

The ICFM is a combined transportation network model that simultaneously solves the user equilibrium route choice model, mode choice model, and trip distribution model. It estimates interregional commodity flows on highway and railway networks by having an Input-Output model as a constraint (Kim, Ham, and Boyce 2002; Ham, Kim, and Boyce 2005a, 2005b; Sohn et al. 2002; Sohn et al. 2003).

The FDLM estimates final demand loss due to disruption of the network. It is based on a resiliency factor, which represents how resilient the production of a certain economic sector is to a disruption of network links that disrupts the commodity flows for a zone (Sohn et al. 2002; Sohn et al. 2003). Note that both the ICFM and FDLM are static models based on a one year period, so the results from the models cannot capture the dynamics of spatio-temporal changes within the analysis period.

The static models can be utilized in new temporal configurations since 3 months (a quarter of a year) is still a valid time interval for long-term models like ICFM and FDLM. However, temporal dependency or interaction cannot be ignored since the damages to the network and the final demand are different among quarters. In other words, the damages to the network will be recovered over time and the decrease of the final demand will be recovered over time, too.

In order to include the temporal changes, ICFM and FDLM are extended to the snapshot ICFM and FDLM, and MRIO is extended to Sequential Inter-industry Model (SIM). The snapshot mathematical models for ICFM and FDLM are suggested as following: (Note that the snapshot ICFM and FDFM are simply adding a time variable into the mathematical model).

Snapshot ICFM

$$\min_{\mathbf{h},\mathbf{x}} \quad Z(\mathbf{h},\mathbf{x}) = \sum_{aw} \int_0^{f_{at}^w} d_{at}^w(\omega)d\omega + \sum_{mjw}\left(\frac{x_{jjt}^{mw}}{g^m}\right)d_{jjt}^w + \sum_m \frac{1}{\alpha^m g^m}\sum_{ijw} x_{ijt}^{mw}\ln\left(\frac{x_{ijt}^{mw}}{x_{ijt}^m}\right) + \sum_m \frac{1}{\beta^m g^m}\sum_{ij} x_{ijt}^m \ln\left(\frac{x_{ijt}^m}{\overline{X}_{it}^m}\right) \tag{4.1}$$

Subject to:

$$\sum_i x_{ijt}^m = \sum_n a_j^{mn}\sum_k x_{jkt}^n + y_{jt}^m \quad \text{for all } m \text{ and } j \text{ at time } t \tag{4.2}$$

$$\sum_w x_{ijt}^{mw} = x_{ijt}^m \quad \text{for all } m, i, \text{ and } j \text{ at time } t \tag{4.3}$$

$$\sum_r h_{ijrt}^{mw} = \frac{x_{ijt}^{mw}}{g^m} \quad \text{for all } m, w, i, \text{ and } j \text{ at time } t \tag{4.4}$$

$$h_{ijrt}^{mw} \geq 0 \quad \text{for all } m, r, w, i, \text{ and } j \text{ at time } t \tag{4.5}$$

where the exogenous variables are:

a^{mn} = technical input-output coefficient representing the inputs from sector m; required to make one unit of output of sector n;

α^m = cost sensitivity parameter for sector m;

β^m = cost sensitivity parameter for sector m;

d_{jj}^w = intraregional distance for region j by mode w (miles);

g^m = factor for converting sector m from dollars to tons ($/ton);

$\overline{X}_i^m$ = total estimated output of sector m in region i ($); and

y_j^m = final demand (consumption, investment and government expenditures) for sector m in region j ($)

The endogenous variables are:

$d_a^w(f_a^w)$ = distance function of total flow on link a by mode w (miles);

f_a^w = total flow on link a by mode w (tons) = $\sum_m \sum_{ijr} h_{ijr}^{mw} \varphi_{ijr}^{aw}$;

φ_{ijr}^{aw} = 1 if link a belongs to route r from region i to region j by mode w, and 0 otherwise;

h_{ijr}^{mw} = flow of output of sector m from region i to region j on route r by mode w (tons);

x_{ij}^m = flow of the output of sector m from region i to region j ($);

x_{ij}^{mw} = flow of the output of sector m from region i to region j by mode w ($);

h = vector of route flows; and

x = vector of interregional flows

Snapshot FDLM

$$\Delta f_t = (I - A)\{[D_t \otimes (1_{13} - R)] \circ [(I - A)^{-1} f_t]\} \tag{4.6}$$

where

Δf_t = change of the final demand by sector by zone at time t;

A = 13 sector by sector direct input coefficient by zone;

D_t = network disruption ratio by zone at time t;

1_{13}^T = $(1 \cdots 1)$;

$1_{13}^T - R$ = one minus sectoral resiliency factor vector;

f_t = final demand by sector by zone before the earthquake at time t;

$\otimes$ = tensor; and

$\circ$ = defined as $B \circ G = (b_{ij} \times g_{ij})_{m \times n}$ where $B = (b_{ij})_{m \times n}$ and $G = (g_{ij})_{m \times n}$

Based on these models, two types of STAMs are proposed: 1) STAM before unscheduled event and 2) STAM after unscheduled event. Before describing these two models, the quarterly Sequential Interindustry Model (SIM) is explained because it provides STAMs with dynamics of economies.

4.4.1 Sequential Inter-industry Model (SIM)

In order to capture temporal economic characteristics and temporal dependency, a static MRIO model needs to be extended to the Sequential Interindustry Model (SIM), which is utilized in the framework of STAMs. Romanoff and Levine (1981) first introduced the SIM, which is an extension of input-output framework. With fixed time intervals, the SIM extends the static input-output model to make it more dynamic.

Table 4.5. Production mode classification

Production Mode	Sector ID	Industries (Sectors)
Long Anticipatory	1	Agriculture, Forestry, and Fisheries
	2	Mining
Short Anticipatory	6	Primary Metals Industries
	7	Fabricated Metal Products
	8	Industrial Machinery and Equipment
	9	Electronic and Electric Equipment
	10	Transportation Equipment
	12	Other Durable Manufacturing
Responsive	3	Construction
Just-in-time	4	Food and Kindred Products
	5	Chemicals and Allied Products
	11	Other Non-Durable Manufacturing
	13	TCU*, Services, and Government enterprises

* Transportation, Communication, and Utilities

A SIM for STAM is developed for a quarter of a year (3 months) by adopting the model developed by Okuyama, Hewings and Sonis (2004). The SIM has different production modes depending on the characteristics of the sectors. The 13 sectors are assigned to four production modes: long anticipatory mode, short anticipatory mode, responsive mode, and just-in-time mode. The sectors in long anticipatory mode anticipate demand four quarters (one year) ahead while the sectors in short anticipatory mode anticipate demand one quarter (3 months) ahead. The sectors in responsive mode respond to the demand one quarter ago. The sectors in just-in-time mode adjust production depending on current demand. Detailed classifica-

tion is shown in Table 4.5. Note that classification of industries will be explained in section 4.5.1.

SIM assumes that the production structure, represented as technical coefficient (A), is constant over time, there are no inventory and capacity limitations, and the future final demands are perfectly predictable.

Based on the production modes and the assumption, the following formulation suggested by Okuyama, Hewings and Sonis is used for the quarterly SIM (2004, pp. 100-101):

$$x_t = A_{al}x_{t+4} + A_{as}x_{t+1} + A_r x_{t-1} + A_j x_t + y_t \tag{4.7}$$

where,

x_t = total output at time t

y_t = final demand at time t

A_{al} = MRIO technical coefficient matrix for the long anticipatory sectors with 4 quarters (one year) anticipation

A_{as} = MRIO technical coefficient matrix for the short anticipatory sectors with 1 quarter anticipation

A_r = MRIO technical coefficient matrix for the responsive sectors with 1 quarter response period

A_j = MRIO technical coefficient matrix for the just-in-time sectors

And, solving the above system yields the following equation (Okuyama, Hewings and Sonis 2004, pp. 100-101):

$$x_t = \sum_{k=1}^{\infty} A_a^k y_{t+k} + \sum_{k=1}^{\infty} A_r^k y_{t-k} + \sum_{k=0}^{\infty} A_j^k y_t + \sum_{k=-\infty}^{\infty} G_k(A_{al}, A_{as}, A_r, A_j) y_{t-k} \tag{4.8}$$

where,

A^k = k th power of MRIO technical coefficient matrix

$G_k(A_a,A_r,A_j)$ = a matrix function whose ij element contains the sum of synergetic path gains among different production modes from industry i to j with a total delay of k.

From SIM, the quarterly total output (x_t) will be calculated and then utilized by STAMs as an important input. The relationship between SIM and STAMs will be described in next section.

4.4.2 Spatio-temporal Analysis Model for a Priori Unscheduled Event (STAM-1)

Before/without the unscheduled event, snapshot ICFM and quarterly SIM provides the interregional and intraregional commodity flow, as well as the commodity flow on each link with two modes, such as highway and railway. By using quarterly SIM, each quarterly snapshot ICFM reflects the quarterly characteristics of the economy.

This model needs the temporal constraint. In other words, decision-makers dictate the timeframe analyzed by the model. For example, if the user wants to know the commodity flow on the transport network for 3 quarters starting with the 1st quarter of 2005, STAM-1 produces analysis for 1st quarter, 2nd quarter, and 3rd quarter of 2005.

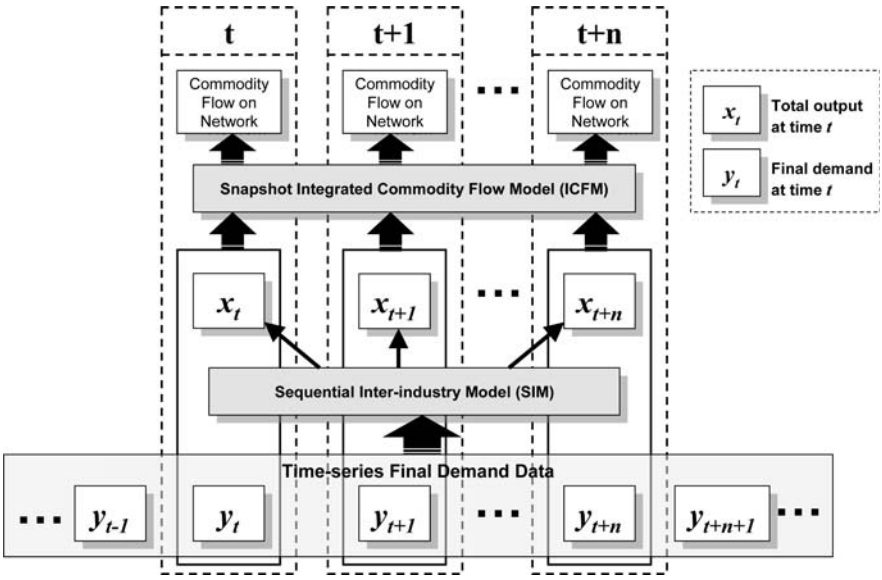


Fig. 4.3. Spatio-temporal analysis model for a priori unscheduled event (STAM-1)

With the temporal constraint, Figure 4.3 shows STAM-1. The snapshot ICFM at a certain quarter (t) needs the final demand at t (y_t) and total output at t (x_t) as inputs. The final demand at t is exogenous. The total output

at t is estimated by quarterly SIM described in section 4.4.1. When quarterly SIM estimates the total output, it needs time-series data of final demand, including before t and after t. Since the total output for snapshot ICFM reflects the temporal dependency of the economy by quarterly SIM, each snapshot ICFM also reflects the temporal characteristics of the national economy.

4.4.3 Spatio-temporal Analysis Model for a Posteriori Unscheduled Event (STAM-2)

After the unscheduled event, such as an earthquake, the status of the national economy begins to change dramatically because the transportation network is disrupted and the disruption of the network will have to be recovered through time. Network disruption has two impacts on the economy. First, the pattern of commodity flow changes, since the commodity flow will reroute due to increase in the travel cost of certain links. Secondly, final demand will be decreased because the input flow to production will decrease. These impacts at time t will cause the changes at time t+1 and so on.

These spatio-temporal interactions are established in Spatio-temporal Analysis Model 2 (STAM-2) as shown in Figure 4.4. Note that the impact at time t feeds into the status of time t+1 via Most Likely Path Flow (MLPF), letting us have the path flow of the commodity. In other words, it depicts which flows pass through the zone and which flows start and stop at the zone. This information is utilized when the zonal disruption ratio is calculated. The flows which pass through at a certain zone are discarded when calculating the zonal disruption ratio.

The unscheduled event (e.g. earthquake) occurring at time t causes network damage at time t. The disruption at t changes the zonal disruption ratio and it is fed into FDLM. With the disrupted network and changed final demand determined by FDLM, ICFM assigns the commodity flow on the highway and railway networks at time t. From the result of ICFM at time t, MLPF is generated. It changes the zonal disruption ratio at time t+1 with the network disruption. Note that the network disruption at time t+1 might be different from the network disruption at time t since the damages are being recovered gradually. The interactions last until t+n which the decision-maker or user designates. In other words, if the user would like to explore the temporal changes in the economy after n time intervals (at t+n) since the earthquake, the iterations might last until time t+n. Or, this model iterates the interaction until the economy is recovered fully. By using this analysis model, the economic network loss over time can be captured and analyzed.

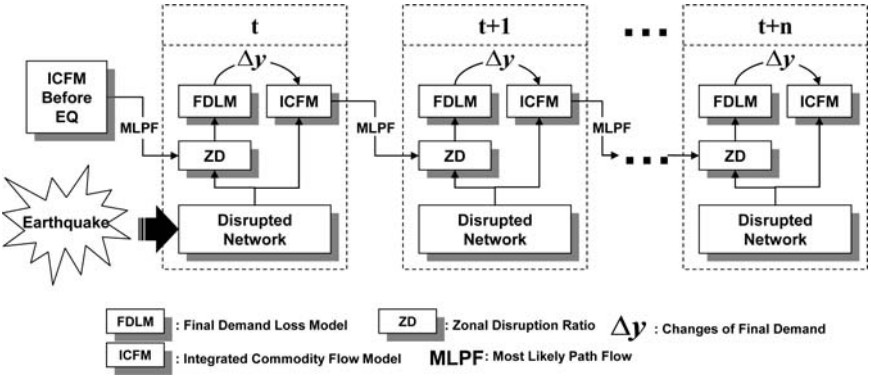


Fig. 4.4. Spatio-temporal analysis model for a posteriori unscheduled event (STAM-2)

4.5 Data and Preliminary Research Results

In this section, the data we have for the research and the data needed for further research are described. In addition, preliminary research results are presented to illustrate how the research can be implemented. Note that some of results are excerpted from previously published materials.

4.5.1 Data for the Research

According to previous research done by by Kim, Ham, and Boyce (2002), Ham, Kim, and Boyce (2005a, 2005b), Sohn et al. (2002) and Sohn et al. (2003), the data for this research consists of four parts: analysis zones, transportation networks, bridges data and economic data. The difference from the previous work lies in the fact that this research requires time-series economic data such as quarterly final demands.

First, the analysis zones are defined and termed EQAZ (Earthquake Analysis Zones) in the previous work (Kim, Ham, and Boyce 2002; Ham 2001; Ham, Kim, and Boyce 2005a, 2005b; Sohn et al. 2002; Sohn et al. 2003). Two sets of zones are developed consisting of 36 zones (shown in Figure 4.5) and 83 zones (shown in Figure 4.6) respectively. These sets of zones are constructed based on county boundaries, state boundaries and National Transportation Analysis Regions (NTAR). Thus, 36 EQAZs are defined and used to analyze the mid-west states in detail in terms of the commodity flow; 83 EQAZs are an expanded version of the 36 EQAZs in order to take the other states into account.

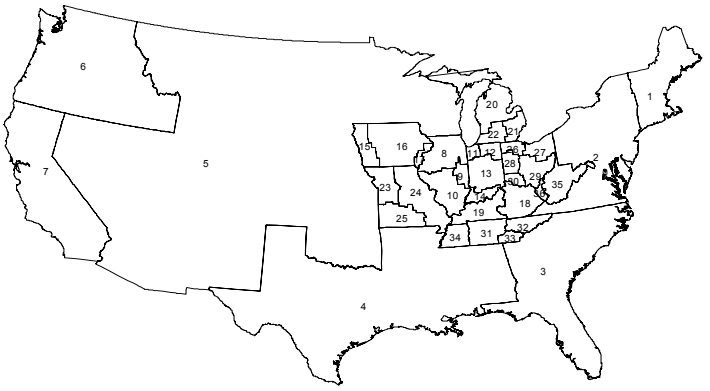


Fig. 4.5. Thirty-six earthquake analysis zones

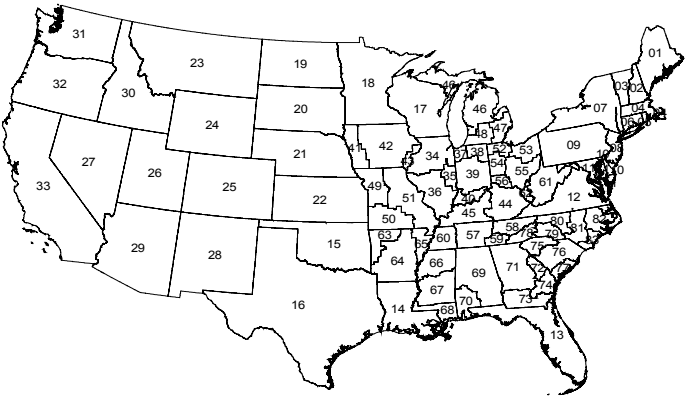


Fig. 4.6. Eighty-three earthquake analysis zone

Secondly, the transportation network data has two modes: highway (shown in Figure 4.7) and railway (shown in Figure 4.8). The highway network is constructed based on the National Highway Planning Network (NHPN), with the interstate highway network extracted using GIS. The major strength of the NHPN is that it contains the linear referencing system information. The railway network is based on the railway layer in the National Transportation Atlas Database (NTAD) from 2003; heavy traffic railways were used to construct the railway network.

Fig. 4.7. Highway network

Fig. 4.8. Railway network

The third component of the data are the bridge data. These data were developed from the National Bridge Inventory (NBI) for 2002, which is comprehensive bridge data collected by the Federal Highway Administration (FHWA). The format of the NBI is a plain ASCII file in table shape. In order to use the NBI data for the research, two tasks are done. First, the ASCII data are converted into MS Access database format to facilitate easier use of the data. Second, bridges on the network are located since the location data in the NBI, such as longitude and latitude, is hidden because of national security reasons. The linear referencing systems in NBI and NHPN are utilized to locate and match the bridges on the network. How-

ever, some states do not have any matched bridges because of wrong or missing linear referenced data.

The final data needed are the economic data for the Multi-regional Input Output Model (MRIO). The data for MRIO consists of MRIO coefficients, trade coefficient and final demand. These are classified into 13 sectors (or industries) as shown in Table 4.6. Note that these are available for 83 and 36 EQAZs.

MRIO coefficients are based on the national input-output (I-O) table published by the United States Bureau of Economic Analysis (BEA). Trade coefficients are calculated based on the Commodity Flow Survey (CFS) of 1997, published by US Census Bureau and Bureau of Transportation Statistics (BTS). Final demands are available annually and quarterly from year 1997 to year 2022. These are estimated based on the Regional Econometric Input-Output Model (REIM) developed by the Regional Economics Applications Laboratory (REAL) at the University of Illinois at Urbana-Champaign and the Federal Reserve Bank of Chicago.

Table 4.6. Sectoral classification

Sector ID	Sectors
1	Agriculture, Forestry, and Fisheries
2	Mining
3	Construction
4	Food and Kindred Products
5	Chemicals and Allied Products
6	Primary Metals Industries
7	Fabricated Metal Products
8	Industrial Machinery and Equipment
9	Electronic and Electric Equipment
10	Transportation Equipment
11	Other Non-Durable Manufacturing
12	Other Durable Manufacturing
13	TCU*, Services, and Government Enterprises

* Transportation, Communication, and Utilities

4.5.2 Preliminary Research Results of Economic Network Loss Analysis

Feasibility of the proposed research can be expressed through the preliminary research results. The results of the economic network loss analysis are presented to show how the Integrated Commodity Flow Model (ICFM) can be used in post-earthquake analysis. Based on the data described in section 4.5.1, the economic network loss is calculated under an unscheduled event by using ICFM (Kim, Ham, and Boyce 2002; Ham,

Kim, and Boyce 2005a, 2005b; Sohn et al. 2002; Sohn et al. 2003). The analysis shows how the economic impacts of the links differ from common expectations.

Figure 4.9 presents five scenarios from analysis of the ICFM reported in Ham's dissertation (2001). Scenario A is total disruption of the links on I-94 between Chicago, IL and Gary, IN. Scenario B disrupts the links on I-65 between Louisville, KY and Nashville, TN. Scenario C disrupts the links on I-40 between Little Rock, AR and Nashville, TN. Scenario D is the combination of B and C, and scenario E is the combination of A, B, and C. According to the results of the study, disruption of I-94 in scenario A has a greater impact on economic activities than disruption of I-65 in scenario B and I-40 in scenario C, even though the location of scenario A is further from the New Madrid fault epicenter than scenarios B and C.

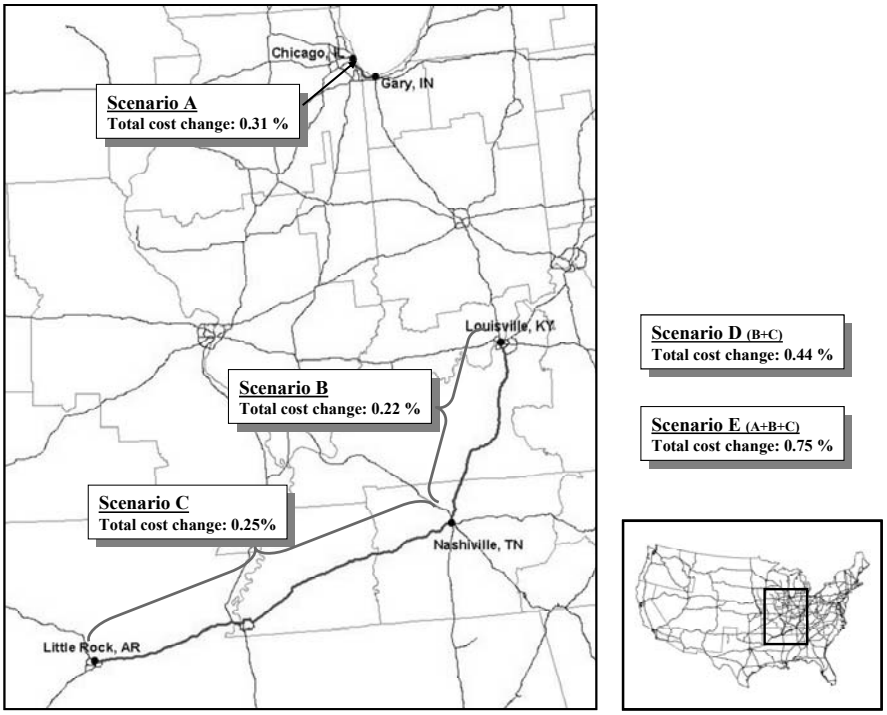


Fig. 4.9. Result of scenario analysis of ICFM (Ham 2001, p. 117)

The commodity flow changes are shown in Table 4.7, illustrating the changes of commodity flow in terms of transportation modes and types of flow. In scenario A, interregional commodity flows equaling as much as $7.21 billion are transferred to the intraregional commodity flows because

of increased shipping costs for highways due to the disruption of the highway network section. In addition, the total commodity flows for Origin-Destination (OD) pairs via highway mode are transferred, by as much as $3.13 billion, to railway mode.

Table 4.7. Result of scenario A: changes of commodity flows (Ham 2001, p. 118)

Scenario	Commodity Flow	Without Event	With Event	Difference	Rate (%)
A	Interregional	2863.80	2856.59	-7.21	-0.25
	Intraregional	1061.19	1068.40	7.21	0.68
	Total Highway OD	3064.95	3061.82	-3.13	-0.10
	Total Railway OD	860.04	863.17	3.13	0.36

As shown in this result, capturing the economic impact or significance of highway or railway links is possible using ICFM. Note that the analysis is based on the yearly based static model.

4.6 Implementation Plan

The multi-regional input output analysis for each quarter has been estimated by SIM (Sequential Inter-industry Model) from year 1997 to year 2016. In addition, the technology to implement the spatio-temporal database has been reviewed and tested.

The next step of this research is refining the spatio-temporal data model from conceptual design to physical design incorporating ISO standards. In other words, with the spatio-temporal data model suggested and the mathematical model, the database for network loss analysis needs to be implemented.

PostgreSQL 8.0.3 will be used as the DataBase Management System (DBMS). In order to implement the spatial features, open source software including PostGIS 1.0.3, a spatial extension to PostgreSQL, will be used.

An additional concern with designing a database is facilitating exploratory analysis on the results of network loss analysis utilizing Geographical Information Systems (GIS). Shaw and Xin (2003) mentioned that temporal GIS or spatio-temporal databases are helpful and critical components for exploratory analysis about spatio-temporal changes, such as the interaction between land use and transportation. Therefore, the database needs to be designed to support this type of exploratory analysis.

Since implementation of the spatio-temporal data model and analysis model should be tested, scenario analysis on the models will be performed with historical earthquakes such as the New Madrid earthquake (circa

1811, Moment Magnitude: 8) and the Northridge earthquake (circa 1994, Moment Magnitude: 6.7).

Acknowledgements

This project on which this chapter was based was supported primarily by the Earthquake Engineering Research Centers Program of the National Science Foundation under Award Number EEC-9701785 to the Mid American Earthquake Center at the University of Illinois at Urbana-Champaign.

References

Abraham, T., and Roddick, J. F. (1999). Survey of Spatio-Temporal Databases. GeoInformatica, 3 (1), pp. 61-99.

Armstrong, M. P. (1988). Temporality in Spatial Databases. In Proceedings of GIS/LIS '88, Vol. 2. (San Antonio, TX), pp. 880-889.

Claramunt, C. and Thériault, M. (1995). Managing Time in GIS: An event-oriented approach. Recent Advances on Temporal Databases, Clifford, J. and Tuzhilin, A. (eds.), Springer-Verlag, Zurich, Switzerland, pp. 23-42.

El-Geresy, B. A., Abdelmoty, A. I., and Jones, C. B. (2002). Spatio-Temporal Geographic Information Systems: A Causal Perspective. In Manolopoulos, Y. and Navrat, P., Lecture Notes in Computer Science, 2435, pp. 191-203.

Hägerstrand, T. (1970). What about People in Regional Science?. Papers of the Regional Science Association. 24, pp.7-21.

Ham, H. (2001). An Integrated Interregional Input-Output and Transportation Network Model for Assessing Economic Impacts of Unexpected Events (Doctoral dissertation, University of Illinois at Urbana-Champaign, 2001). DAI-A 62/06, p. 2255

Ham, H., Kim, T. J., and Boyce, D. E. (2005a). Implementation and Estimation of a Combined Model of Interregional, Multimodal Commodity Shipments and Transportation Network Flows, Transportation Research B, 39(1), pp. 65-79

Ham, H., Kim, T. J., and Boyce, D. E. (2005b). Assessment of Economic Impacts from Unexpected Events Using an Interregional Commodity Flow and Multimodal Transportation Network Model. Transportation Research A, 39(10), pp. 849-860

Kim, T. J., Ham, H., and Boyce, D. E. (2002). Economic Impacts of transportation network changes: Implementation of a combined transportation network and input-output model, Papers in Regional Science, 81, pp. 223-246. Springer-Verlag, Berlin.

Langran, G. (1988). Temporal GIS design tradeoffs. In Proceedings of GIS/LIS '88, Vol. 2. (San Antonio, TX), pp. 890-899.

Langran, G. (1992). Time in Geographic Information Systems. Bristol, PA: Taylor & Francis Inc.

Okuyama, Y., Hewings, G. J. D., and Sonis, M. (2004). Measuring Economic Impacts of Disasters: Interregional Input-Output Analysis Using Sequential Interindustry Model. In Okuyama, Y., and Chang, S.E. (Eds.), Modeling Spatial and Economic Impacts of Disasters. Springer

Pang, M. Y. and Shi, W. (2002). Development of a Process-Based Model for Dynamic Interaction in Spatio-Temporal GIS. GeoInformatica, 6(4), pp. 323-344.

Peuquet, D. J. (1994). It's about Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems. Annals of the Association of American Geographers, 84(3), pp. 441-461.

Peuquet, D. J. (2001). Making Space for Time: Issues in Space-Time Data Representation. GeoInformatica, 5(1), pp. 11-32.

Peuquet, D. J., and Duan, N. (1995). An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. International Journal of Geographical Information Systems, 9(1), pp. 7-24.

Romanoff, E. and Levine, S.H. (1977). Interregional Sequential Interindustry Modeling: A Preliminary Analysis of Regional Growth and Decline in a Two Region Case. Northeast Regional Science Review, 7, pp.87-101.

Shaw, S.-L and Xin, X. (2003). Integrated land use and transportation interaction: a temporal GIS exploratory data analysis approach. Journal of Transport Geography, 11, pp. 103-115

Sohn, J., Hewings, G. J.D., Kim, T. J., Lee, J. S., and Jang, S.-G. (2002) Analysis of Economic Impacts of an Earthquake on Transportation Network. Available at http://epil.urban.uiuc.edu/publications/02-0801.pdf.

Sohn, J., Kim, T. J., Hewings, G. J.D., Lee, J. S., and Jang, S.-G. (2003). Retrofit Priority of Transport Network Links under an Earthquake, ACSE Journal of Urban Planning and Development, 129-4, pp. 195-210.

Worboys, M. F. (1994a). A Unified Model for Spatial and Temporal Information. The Computer Journal, 37, pp. 26-34.

Worboys, M. F. (1994b). Unifying the Spatial and Temporal Components of Geographical Information. In Waugh, T.C. and Healy, R.G., editors, Advances in GIS Research, Proceeding of the 6th International Symposium on Spatial Data Handling, Taylor & Francis,.

Yuan, M. (1994). Wildfire Conceptual Modeling for Building GIS Space-Time Models. GIS/LIS ’94 , (Phoenix), pp. 860-869.

5 Vulnerability: A Model-Based Case Study of the Road Network in Stockholm

Katja Berdica[1] and Lars-Göran Mattsson[2]

[1] Transek AB, Sundbybergsvägen 1A, SE-173 73, Solna, Sweden; Email: katja.berdica@transek.se
[2] Department of Transport and Economics, Royal Institute of Technology, SE-100 44 Stockholm, Sweden; Email: lgm@infra.kth.se

5.1 The Why and How

5.1.1 Background and Scope of the Study

Vulnerability, exposure and criticality in various infrastructures are issues that have been more explicitly looked into in recent years. However, road vulnerability as such has not been in focus for very long, despite the fundamental importance of our road networks in everyday life, as well as in crisis evacuation situations. Consequently, network reliability in transport modelling is an important and growing field of research (Lam 1999). The connection between reliability, vulnerability and other related concepts are discussed in Berdica (2002), with the main proposition that vulnerability analysis of road networks should be regarded as an overall framework, within which different transport studies can be performed to describe how well our transport systems function when exposed to different kinds of disturbances. Following that approach, this paper presents the results from a model-based case study, performed with the overall objective to study how vulnerable the Stockholm road network is in different respects. More specifically it is built up around three main questions:

1. How do interruptions of different critical links affect the system and how important are these links in relation to one another?
2. How is the network performance affected by general capacity reductions and possible prioritisation of a sub-network?

3. How is the system affected by traffic demand variations, i.e. how close to its capacity limit does the system operate?

How sensitive a (transport) network is to having its links closed down can be studied in different ways. Wollmer (1964) considers the problem of removing *n* links in a capacitated network such that the reduction in the maximum flow between an origin and a destination is maximised. Although this approach may be relevant in many situations, the focus of the present study is on travel time rather than on maximum flow. Therefore, the approach by Ball et al. (1989) is more relevant. They consider an uncapacitated network for which there is a distance (or travel time) and a removal cost associated with each link. They then pose the problem of removing *n* links such that the increase in the shortest distance (travel time) between an origin and a destination is maximised given a budget for the removal of links. Jenelius et al. (2006) apply similar ideas in a recent vulnerability study of the uncongested, undirected network of northern Sweden (see also Taylor and D'Este 2004). An origin-destination trip (OD) matrix is given for the network. One link at a time is removed from the network and then replaced. For each such removal, some trips cannot reach their destinations (i.e. yielding infinite travel times) or there is a finite increase in total travel time, as the trips that used the removed link are rerouted to find their new fastest routes. Results are presented for all trips as well as for the trips starting from each separate municipality in the study area. As in the approach by Ball et al. (1989), these calculations are carried out under the assumption that link travel times are unaffected by the changes in the loads due to the rerouting of the trips. This is quite reasonable for a rural network with almost no congestion. For the congested Stockholm network such an assumption is not valid. For this case it is necessary to apply a network assignment model that allows link travel times to vary with link loads. This increases the computational burden significantly and it is necessary to restrict the analysis to a handful of carefully selected scenarios.

The network considered in the present analysis is that of the entire Stockholm Region (see Fig. 5.1). The area has a total of about 1.8 million inhabitants and stretches over 6500 km^2.

This study focuses on disturbances of such duration that it is reasonable to assume that the road users are informed of the network status and have changed their behaviour accordingly. There are a number of behavioural mechanisms through which travellers adapt to new network conditions. For a disturbance lasting a couple of days or some weeks, which is assumed here, changing trip frequency, travelling at another time of day, choosing another route or another mode of transport would be considered.

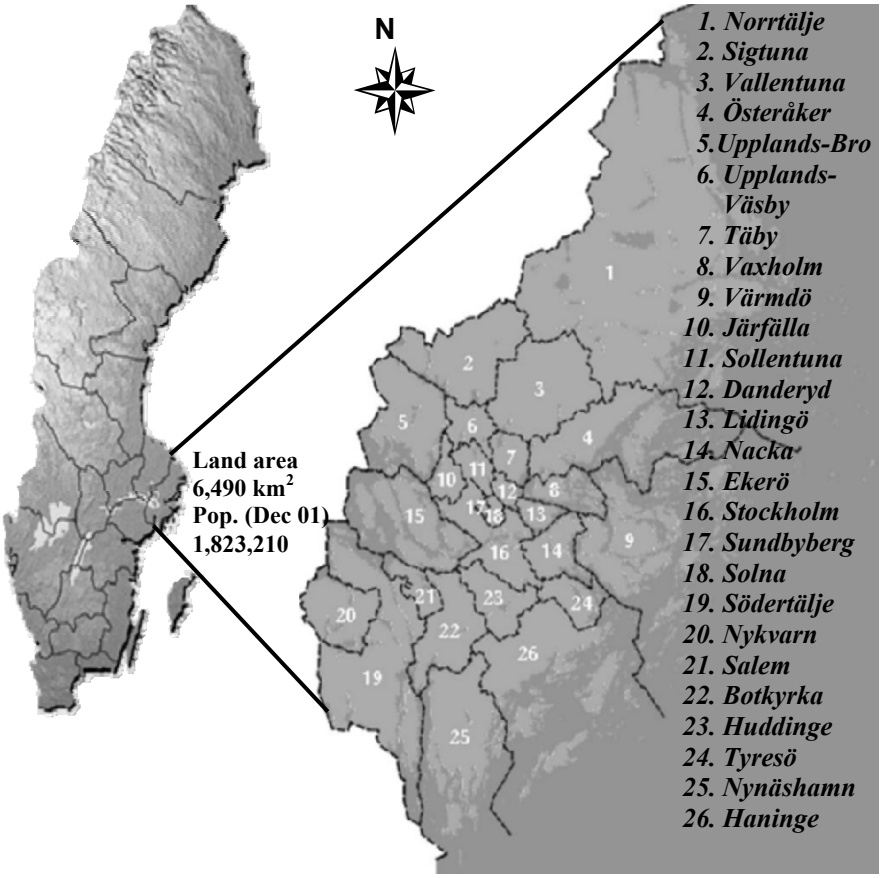


Fig. 5.1. The Stockholm Region with its 26 municipalities

Changing one's destination would be less plausible although it could be an alternative for certain types of errands. Longer-term changes such as changing car-ownership and reallocation of home and/or place of work would hardly be considered in the present context. In order to simplify the analysis and to highlight the results in a more distinct way, the present study considers adaptations in route choice under equilibrium conditions only and the effect that this has on travel times and distances for the road users in the Stockholm Region. It hence concerns a fixed car travel demand, not allowing for changes in time period, trip frequency and/or transfers to other modes. This may seem inappropriate, since not allowing for

possible reductions in car traffic could lead to an overestimation of the congestion problems after an incident. However, studies show that people in general do not abandon their private car so easily (Transek 1999). Very long delays must be the case before most people would consider choosing to go by public transport, even when information on the disturbance is available beforehand. Therefore, excluding by assumption transfers to public transport was deemed acceptable for the present case study. To exclude changes in trip timing and hence peak spreading is more questionable. This was still necessary because the available regional model for Stockholm does not include such behavioural mechanisms.

The volume-delay (vd) functions are hence fundamental for the representation of the behavioural responses to the disturbances in the network. They express travel time t (minutes) on each link as a function of traffic volume V (vehicles per hour and lane) and link length l (kilometres). Their general construction (as presented below) is a polynomial curve, combined with a linear function beyond the reference capacity $\overline{V}$ (vehicles per hour and lane), such that the function but not its derivative is continuous:

$$t(V,l) = \left[\frac{V}{k_1} + k_2 \left(1 + \left(\frac{V}{k_3} \right)^n \right) \right] \times l + k_4 \quad \text{if } V \le \overline{V} \tag{1}$$

$$t(V,l) = t(\overline{V},l) + k_5 (V - \overline{V}) \quad \text{if } V \ge \overline{V} \tag{2}$$

where k_i are speed limit and traffic condition dependent parameters and n is the "polynomial" parameter.

The functions are different for different links depending on their respective speed limit and traffic conditions, e.g. traffic is deemed more "disturbed" in the city centre than on peripheral links. Because of the link length dependency, the exact form of the corresponding graphs will vary accordingly. The three vd-functions used on 50 km/h links for unit length of 1 km are shown in Fig. 5.2, while an explicit listing of all vd-functions and applied parameter values can be found in Berdica (2000).

As traffic volume approaches the reference capacity, link travel time should increase rapidly. For final equilibrium solutions on the left hand side of this point, the form of the curve to the right does not matter. It simply ensures the convergence of the algorithm by which the equilibrium solution is found. For solutions to the right, the linear function is designed to impose an extra travel time (independent of link length, as opposed to the polynomial part that is length dependent), to take into account that exceed-

ing the reference capacity will result in queues.[1] To represent this queuing time in a realistic way is very tricky, however, since it is very much dependent on the length of the period for which this "overloading" lasts.

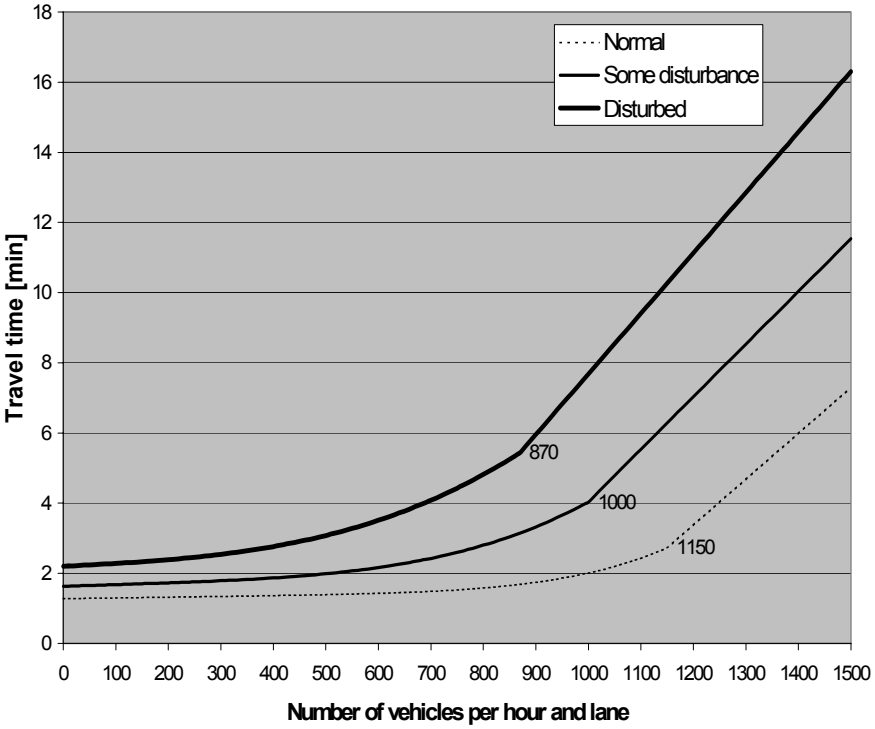


Fig. 5.2. Vd-functions used in EMME/2 for links of length 1 km and speed limit 50 km/h in various locations in the network. The numbers in the graph indicate reference capacity $\overline{V}$ according to eqs. (1) and (2). The corresponding parameter vector **k** = (1267, 2, 802, 0.2, 0.017), (2083, 1.4283, 928, 0.2, 0.015) and (4743, 1.2767, 1162, 0, 0.013), and n = 3, 4 and 6, for the vd-functions from top to bottom, respectively

The most important implication of choosing EMME/2 for this case study comes from the construction of the vd-functions and the application of the principle of user equilibrium: the modelled situation is that of a steady-state, but the linear part of the vd-function beyond reference capacity means that there is no actual capacity limit – all traffic being let

[1] The basis is the theoretical "clearing time" for a queue built up during one hour. In theory, the average queuing time is half of this value. The extra time added (i.e. "real" queuing time) is then again half of the theoretical queuing time, based on an assumption that the average bottleneck consists of two consecutive links.

through is just a question of time. The resulting traffic volumes during e.g. the morning peak hour are hence to be regarded as *manifested* demand for travel on a link, rather than *actual* traffic flows during that hour. Also, the presently used algorithm is based on link flows only, and the feature of queues spilling over backwards on adjacent links is not captured. At the same time links downstream from an overloaded link are experiencing the same "excess" traffic volume (and the associated extra queuing time), although this does not occur in reality since the vehicles in question simply cannot come through. Thereby a secondary purpose of the study crystallises, and that is to investigate to what extent the present traffic equilibrium model is suitable for this purpose. To shed some light upon this, a comparative study was performed using an alternative network equilibrium model that was under development at the time, which uses a different approach as far as solution algorithm and vd-functions are concerned. An alternative approach would have been to use some modern micro simulation model. This may be possible in the future. Implementation of such models is under way, though only for part of the network considered in this study.

5.1.3 Building Scenarios

In addition to a Base Scenario, in which today's traffic (base year 1997) is assigned to the network without "interference", the study includes a total of twelve scenarios as listed in Table 5.1.

Table 5.1. Description of case study scenarios

Scenario	Denomination	Description
1	Essinge Route 1	One lane closed, northbound direction
2	Essinge Route 2	Complete closure, northbound direction
3	Essinge Route 3	Complete closure, both directions
4	Central Bridge	Complete closure, northbound direction
5	Western Bridge	Complete closure, northbound direction
6	Traneberg Bridge	Complete closure, both directions
7	Danvikstull Bridge	Complete closure, both directions
8	General red. in capacity	Reduced "free flow speed", all links
9	Selective red. in capacity	Reduced "free flow speed", secondary road network
10	Car traffic 1	Today's traffic -8%
11	Car traffic 2	Today's traffic +8%
12	Car traffic 3	Today's traffic +16%

The Stockholm road network stretches over a number of islands and the function of the system depends very much on the connecting bridges being passable. From this structure it is easy to realise that some of the most crucial spots in Stockholm are the passages over the waters of Saltsjö-

Mälaren. Therefore these five main bridges were chosen for the study of critical links (Fig. 5.3; reference capacities according to Table 5.2). It is interesting to note that about 14% (485 per year) of the Stockholm Road Assistance Service commissions concern these very bridges. Closures of one or more lanes due to e.g. traffic accidents or physical failure were modelled by simply putting the respective link capacity equal to zero (Scenarios 1–7).

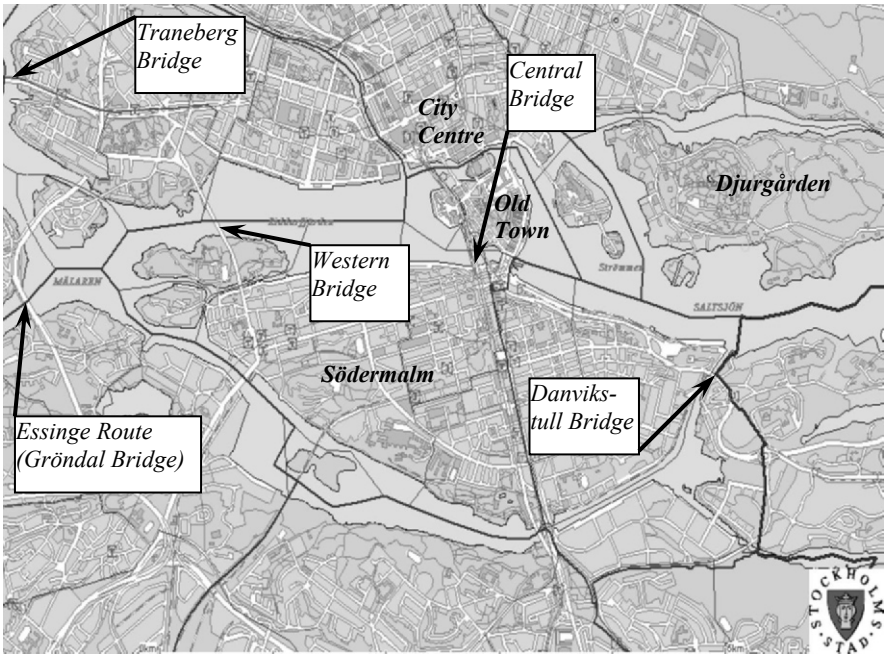


Fig. 5.3. Chosen critical links over Saltsjö-Mälaren, Stockholm

Considering the Swedish climate, one of the most common reasons for general capacity reductions in the road network is snow and/or sleet. According to Swedish Meteorological and Hydrological Institute (SMHI) statistics, days with snow additions of 20 mm or more come about approximately 15 times per season (November–April; Table 5.3). General reductions in road network capacity due to winter weather were modelled by altering the free flow speed in, and the form of, the link vd-functions. The modifications assume a 15% reduction in vehicle speed, based upon the results from studies at the Swedish National Road and Transport Research Institute (VTI) of how different road weather conditions influence

e.g. driving speed (Wallman 1996). Scenario 8 can be said to represent slippery roads in general, while Scenario 9 simulates snow-clearing prioritisation in part of the network. The same fixed OD-matrix was used in both scenarios, since the VTI studies concluded that the presence of snow/sleet did not influence traffic demand noticeably. One would suspect that extremely bad overall weather has a greater effect, but that is not within the scope of this study.

Table 5.2. Reference capacities and number of lanes for the five critical bridges

Bridge	Direction	Reference Capacity [veh/h/lane]	No of Lanes
Essinge Route	Both	2000	3
Central Bridge	Both	2000	3
Western Bridge	Southbound	2000	2
	Northbound	1550	2
Traneberg Bridge	Both	1550	2
Danvikstull Bridge	Both	1150	2

Table 5.3. Snow and temperature statistics for Stockholm-Bromma meteorological station (Source: SMHI; authors' adaptation)

Statistics Nov–Apr 1961–90	Average	Max	Min
Total amount (cm) of snow per winter	79.2	185	12
No of days with newly fallen snow of			
< 2 cm	9.8	21	2
2–5 cm	9.6	18	2
5–7 cm	2.0	7	0
7–10 cm	1.8	4	0
> 10 cm	1.4	4	0
No of days with temp. passing 0°C [a]	73.3	103	51

[a]At least once during 24 hours

As in most major cities, the seemingly ever-increasing traffic volume in Stockholm is becoming of growing concern. More or less normal travel demand variation around the average could also have a great effect, when the system is operating close to capacity limits. Scenarios 10–12 address this issue by simply decreasing/increasing traffic volumes in the OD-matrix representing the travel demand. The factor of 8% was obtained as the estimated average standard deviation for daily traffic volumes, registered during two months 1998 (weekdays in February and October) at six traffic count sections in Greater Stockholm.

5.2 Analysis Results

5.2.1 Modelling Details

Travel demand is assigned in three different time periods: morning rush hour (7–8 a.m.), evening rush hour (4–5 p.m.), and an estimated typical "middle hour". These have then been weighted together, presenting the results on a 24-hour basis. The principle direction of flow during the morning peak is toward the city centre. This period is therefore chosen for the presentation when relevant (basically in Scenarios 1, 2, 4 and 5).

All model calculations are carried out at the detailed level of 1,246 OD-zones (six of these are areas outside the region itself). For the sake of presentation, calculated results such as travel distances, travel times and travel speeds need to be aggregated. All aggregation runs the risk of concealing spatial variation in the results, an effect that is sometimes referred to as the modifiable areal unit problem. To avoid this problem as far as possible, the aggregated zones must be carefully designed. In the present study the original OD-zones were aggregated to ten zones as listed in Table 5.4 and depicted schematically in Fig. 5.4. The aggregation is based on the following principle: the original OD-zones have been aggregated around the major arterial approach roads in Stockholm, without splitting any municipality apart from Stockholm city, and so that the critical bridges considered in Scenarios 1–7 always will be placed on the border between the aggregated zones. In this way, the risk of evening out interesting spatial variation in the results will be minimised.

Note that the average changes in comparison with the Base Scenario are calculated for all trips in the whole of Stockholm Region, and not only for the trips over the critical passage over Saltsjö-Mälaren. The "maximum change" presented is simply the greatest absolute time/distance/speed change (also expressed as a percentage change) compared to the Base Scenario, as found in the resulting aggregated 10×10 OD-relation matrix. It is hence the OD-relation that is "worst off" for each indicator in each scenario. The speed measure used is simply total distance travelled divided by total time taken for all trips in the respective aggregate.

Finally, a short note should be included regarding the convergence of the model for the different scenarios. It can be noted that out of the twelve scenarios × three time periods = 36 model runs, eight were interrupted by the maximum number of iterations stopping criterion. However, if a 1% relative gap criterion had been used, only three runs would not have con-

Table 5.4. Description of the zone aggregation used in the presentation of the results

Denomination	Included municipalities	No of zones
(0) External areas	Municipalities outside the Stockholm Region	6
(1) Western Stockholm	Part of Stockholm city	88
(2) North inner City	Part of Stockholm city	133
(3) South inner City	Part of Stockholm city	44
(4) Southern Stockholm	Part of Stockholm city	135
(5) North-western region	A) Solna, Sundbyberg B) Sigtuna, Upplands-Väsby, Sollentuna C) Upplands-Bro, Järfälla, Ekerö	250
(6) North-eastern region	D) Danderyd, Lidingö E) Vallentuna, Täby, Österåker, Vaxholm F) Norrtälje	248
(7) Eastern region	G) Nacka H) Värmdö	87
(8) Southern region	I) Tyresö, Haninge J) Nynäshamn	96
(9) Western region	K) Salem, Huddinge, Botkyrka L) Södertälje, Nykvarn	159
Sum		1246

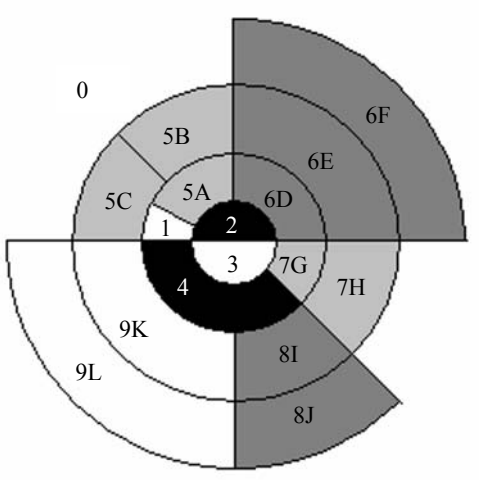


Fig. 5.4. Schematic illustration of the zone aggregation (capital letters refer to municipalities as per Table 5.4)

verged in a proper manner.[2] For Scenarios 2 and 3, though, the final equilibrium solution was not always obtained, the greatest deviation being 2.5% for the afternoon rush hour in Scenario 3. Detailed analysis of the result report shows, however, that the remaining differences are small and very unlikely to influence the general results and conclusions drawn.

5.2.2 Winners and Losers: Net Effects

In the base scenario there are about 111,000 trips during the morning rush hour, and a total of 1.6 million trips per day in the Stockholm Region. Average travel time / trip length / travel speed under "normal" conditions (as calculated by EMME/2) are:

- Morning rush hour: 28 min / 17 km / 37 km/h,
- Daily: 25 min / 14 km / 34 km/h.

It may seem counter intuitive at first that average speed during the morning rush hour is higher, although the level of congestion is greater. This is explained by differences in the composition of trip purposes, which in turn has implications for the destination choices and trip lengths. The overall average effects (i.e. at the regional level) on travel time, trip length, and travel speed in the different scenarios are presented in Tables 5.5 and 5.6. It should be noted that in virtually all scenarios (except the ones where it gets worse or better for everybody – basically Scenarios 8–12) there are both winners and losers, although the former are most often in a minority.

The closure of a specific link will force travellers on routes including that link to choose another route to their destinations. This means that travel demand is also reduced on other links on the abandoned route, leaving them less congested for traffic that is unaffected by the closure. As a second order effect this may attract other traffic to the now less congested links, hence changing congestion on still other links. In the end the travellers on these links will, or will not, experience a travel time reduction. The total effects are intuitively very difficult to foresee but should be captured in a reasonable way by the user equilibrium assumption of the EMME/2 model. Route choice is based upon minimisation of travel time, which means that the chosen route is not necessarily the shortest. Changes in traffic load on certain links may, however, turn a previously shorter but slower

[2] The stopping criterion used is the maximum of 30 iterations, 0.2% relative gap or 0.2 minutes normalised gap, any of which occurs first. Empirically, a relative gap of 1% or less is considered sufficiently close to a perfect equilibrium (INRO 1998, pp. 6–17).

route into the fastest alternative. Hence, the closing of a link can also result in trip distance becoming less in some travel relations. Because of this, the average effect in terms of an absolute figure can sometimes seem insignificant (Tables 5.5 and 5.6). It is then of greater interest to study the effects in specific travel relations etc., which will be done in later sections. Average effect can, however, be quite large when related to the normal situation, as seen from the percentage changes for e.g. Scenario 3.

Table 5.5. Mean effects at the regional level in different scenarios compared to the Base Scenario, calculated on a daily basis

Scenario	Travel time [min/trip]		Trip length [km/trip]		Travel speed [km/h]	
0. Base Scenario	25.0	–	14.2	–	34.1	–
1. Essinge Route 1	-0.1	0%	0.0	–	0.0	–
2. Essinge Route 2	+1.7	+7%	0.0	–	-2.2	-6%
3. Essinge Route 3	+3.7	+15%	0.0	–	-4.5	-13%
4. Central Bridge	+0.6	+2%	+0.1	+1%	-0.6	-2%
5. Western Bridge	+0.1	+1%	0.0	–	-0.3	-1%
6. Traneberg Bridge	+0.5	+2%	+0.2	+1%	-0.4	-1%
7. Danvikstull Bridge	+1.2	+5%	+0.2	+2%	-1.1	-3%
8. General reduction in capacity	+3.5	+14%	0.0	–	-4.3	-13%
9. Selective reduction in capacity	+1.0	+4%	+0.1	+1%	-1.2	-4%
10. Car traffic 1	-0.5	-2%	0.0	–	+0.6	+2%
11. Car traffic 2	+0.5	+2%	0.0	–	-0.7	-2%
12. Car traffic 3	+1.1	+4%	0.0	–	-1.5	-4%

Table 5.6. Mean effects at the regional level in different scenarios compared to the Base Scenario, calculated for the morning rush hour

Scenario	Travel time [min/trip]		Trip length [km/trip]		Travel speed [km/h]	
0. Base Scenario	27.8	–	16.9	–	36.5	–
1. Essinge Route 1	+0.1	+1%	0.0	–	-0.2	0%
2. Essinge Route 2	+4.1	+15%	+0.1	+1%	-4.4	-12%
3. Essinge Route 3	+5.3	+19%	+0.2	+1%	-5.5	-15%
4. Central Bridge	+1.6	+6%	+0.1	+1%	-1.7	-5%
5. Western Bridge	+0.3	+1%	0.0	–	-0.4	-1%
6. Traneberg Bridge	+0.8	+3%	+0.2	+1%	-0.6	-2%
7. Danvikstull Bridge	+1.7	+6%	+0.2	+1%	-1.6	-4%
8. General reduction in capacity	+4.6	+17%	0.0	–	-5.2	-14%
9. Selective reduction in capacity	+1.2	+4%	+0.1	+1%	-1.4	-4%
10. Car traffic 1	-0.7	-3%	0.0	–	+1.0	+3%
11. Car traffic 2	+0.7	+3%	0.0	–	-0.8	-2%
12. Car traffic 3	+1.9	+7%	+0.1	0%	-2.2	-6%

5.2.3 Total Cost Estimation

For an assessment of the effects in economic terms, it is of interest to consider the total costs of extra travel time and/or trip kilometres (compared to the Base Scenario), experienced by all travellers in the Stockholm Region.

Total travel time in the road traffic system in the Base Scenario is just over 660,000 hours per day. Using the official Swedish travel time value of 35 SEK per hour (private regional trips < 100 kilometres in length, 1999) and occupancy rate of 1.46 persons per vehicle, the total value of car travel time under normal conditions is about 34 million SEK per day. On a yearly basis this amounts to 34 million × 250 days = 8.5 billion SEK. The increase in total travel time of 7% on a daily basis in Scenario 2 may seem small. However, expressed in monetary terms (Table 5.7), it means that 2.26 million SEK extra are spent each day that the Essinge Route is closed in the northbound direction. If we were not dealing with a "perfect information" case, this cost would undoubtedly be higher, not only because of a non-equilibrium situation, but also because unforeseen delay is usually assigned a greater discomfort factor and hence a higher cost.

An important clarification is that the travel time differences in the last three scenarios are calculated with respect to the number of trips in the Base Scenario. Hence, the stated effect of e.g. Scenario 12 does not include the extra costs that follow implicitly from a 16% increase in traffic. In other words, if Scenario 12 became a reality, the car travellers of today would face a total extra cost of 1.57 million × 250 days = about 390 million SEK per year.

Table 5.7. Economic implication of change in total travel time (35 SEK/h) and in trip length (1.3 SEK/km) in different scenarios, compared to the Base Scenario, calculated on a daily basis

Scenario	Change in total cost [million SEK]		
	Travel time	Trip length	Sum
1. Essinge Route 1	-0.08	-0.01	-0.09
2. Essinge Route 2	+2.26	+0.02	+2.28
3. Essinge Route 3	+5.04	+0.04	+5.08
4. Central Bridge	+0.74	+0.21	+0.95
5. Western Bridge	+0.20	+0.00	+0.20
6. Traneberg Bridge	+0.70	+0.32	+1.02
7. Danvikstull Bridge	+1.64	+0.49	+2.12
8. General reduction in capacity	+4.76	-0.04	+4.72
9. Selective reduction in capacity	+1.40	+0.21	+1.61
10. Car traffic 1	-0.65	-0.01	-0.67
11. Car traffic 2	+0.63	+0.02	+0.65
12. Car traffic 3	+1.52	+0.05	+1.57

The first scenario apparently gives a slight travel time reduction at the regional level, corresponding to costs decreasing by about 80,000 SEK. This is caused by travel time decreasing in the "middle hour" run of the scenario. Bræss Paradox (see e.g. Sheffi 1985) shows that this is theoretically possible, but it has not been analysed any further whether the effects in this case can be explained by such a phenomenon.

Concerning changes in trip length, the different scenarios cause very little in terms of percentage extra kilometres travelled at the regional level (see Table 5.5). Still, using the perceived marginal cost of 1.3 SEK per kilometre (1999), there is a noticeable extra cost due to this extra distance. The total distance travelled daily in the system under normal conditions is about 22.5 million vehicle kilometres. The greatest change in trip length is an increase of 2% in Scenario 7, when Danvikstull Bridge is impassable. This increase of 374,000 vehicle kilometres per day translates into roughly 0.5 million SEK extra in vehicle operating costs.

For simplicity, the cost per vehicle kilometre used in the estimates above is only concerned with the owners' perceived operating costs. The effect of Scenario 7 is even worse when considering first: the limited number of travel relations affected (to/from the Eastern Zone), and second: a clear concentration of increased traffic volumes on the alternative route in the southeastern part of the region. Assessing these effects at the regional level is hence less appropriate, since the most "damage" is caused along the alternative route that is actually chosen. It is also difficult to put a total price tag on these extra vehicle kilometres. Individuals experiencing increasing vehicle operating costs (apart from the induced increase in travel time cost) is a fairly simple issue to price. Determining the costs for the whole society due to external effects such as increased pollution and noise is more difficult, and further analysis is not within the scope of this study.

5.3 Discussion

In this section, the extensive data material resulting from running the scenarios (some of which have already been presented in previous tables) is analysed and discussed in more detail, with the purpose to highlight the more local effects. A detailed listing of results (matrices and maps) can be found in Berdica (2000).

5.3.1 Saltsjö-Mälaren Passage

The central east-west water strait called the Saltsjö-Mälaren passage divides the Stockholm Region into two halves, with only three major and two minor bridges connecting them (see Fig. 5.3). In Scenarios 2, 4 and 5 these major bridges (links) are each in turn closed in the northbound direction.[3] Since this is the prevailing direction of flow during the morning rush hour, that is the period discussed in the following. Please note that the number of trips across the Saltsjö-Mälaren passage is constant by assumption (Table 5.8).

Table 5.8. Comparison of traffic volumes [vehicles] over the bridges (northbound direction) in Scenarios 2, 4 and 5 during the morning rush hour

Bridge	Base Scenario	Scenario 2: Essinge Route 2	Scenario 4: Central Bridge	Scenario 5: Western Bridge
Bridge at Gröndal	5100	–	+1500	+1000
Central Bridge	4500	+1500	–	+400
Western Bridge	1700	+2500	+1700	–
Other	1500	+1100	+1300	+300

From the changes in traffic volume it is clear that the five bridges are the main alternative routes for each other, in the case of an interruption of such duration that a new equilibrium route choice has been obtained. Since there are only these five bridges ("other" being two small bridges from the Old Town) for crossing the Saltsjö-Mälaren passage, the traffic pressure naturally increases on remaining connections. Closing the Essinge Route (Scenario 2) gives rise to an increase in demand for travel over Western Bridge of almost 150%, resulting in 2050 vehicles per hour and lane, while reference capacity is only 1550. It is then important to remember that all the extra vehicles on e.g. Western Bridge cannot possibly pass during the morning rush hour, and the number of vehicles should be interpreted as the demand for passing, rather than actual traffic flow over the bridge.

To judge from the changes in travel distance in Table 5.9, incidents on Central Bridge give rise to the longest detours when considering maximum change. The average speed reduction is less than half the one resulting from Scenario 2, though. Hence the system has a better capacity for "swallowing" the traffic from Central Bridge, one reason being that the longer

[3] As a matter of fact, on October 14, 2005, a crane ship collided with the Essinge Route so seriously that one of the now four lanes in the northbound direction had to be closed for more than two months. On weekdays during the first two weeks after the collision, the average traffic volume decreased by about 1900 vehicles during the morning rush hour in the northbound direction.

distances involve a greater geographical spread. However, speed is not a very appropriate effect indicator, since longer travel time on a longer route may in fact give a speed increase, thus "concealing" the adverse effect. The relatively speaking small effects of the Western Bridge-scenario is probably explained partly by its location in-between the other two bridges, partly by its lower traffic load in the Base Scenario.

Table 5.9. Comparison of average travel time, trip length and travel speed in Scenarios 2, 4 and 5 during the morning rush hour (changes in relation to Base Scenario)

Information	Scenario 2: Essinge Route 2		Scenario 4: Central Bridge		Scenario 5: Western Bridge	
Travel time						
Mean change [min/trip]	+4.1	+15%	+1.6	+6%	+0.3	+1%
Max. change [min/trip]	+39.9	+91%	+16.6	+37%	+5.4	+13%
Trip length						
Mean change [km/trip]	+0.1	+1%	+0.1	+1%	0.0	–
Max. change [km/trip]	+1.5	+5%	+3.0	+8%	+0.7	+5%
Travel speed						
Mean change [km/h]	-4.4	-12%	-1.7	-5%	-0.4	-1%
Max. change [km/h]	-19.0	-41%	-7.6	-16%	-4.5	-12%

The most affected travel relations are naturally from the zones south of, to the zones north of, the Saltsjö-Mälaren passage. For these, travel time increases with as much as 20–40 minutes per trip in Scenario 2. From this point of view, a closure of the Essinge Route in the northbound direction is the most critical one, as far as consequences are concerned.

5.3.2 Other Critical Bridges

Essinge Route 3, Traneberg Bridge and Danvikstull Bridge are scenarios in which the links are completely cut off in both directions. This comparison is made on a daily basis (Table 5.10). In case of a complete closure of the Essinge Route, again the most crucial travel relations are across the central water strait in the city. The effect in terms of number of relations affected is of course greater, since vehicles in both directions experience the altered conditions. The increase in travel time results in decreased travel speed, since Scenario 3 does not result in significant detours. Only just under 10% of the travel relations experience increased trip distance by half a kilometre or more. As expected, Scenario 6 affects trips to and from Western Stockholm the most, but there are effects in some north-south (and vice versa) relations as well. For Danvikstull Bridge, the travel time increases are without exception concentrated to journeys from and to the Eastern Region, hence the great difference compared to the average regional effects

of Scenario 7. The observed differences between Scenarios 6 and 7 regarding increased trip length are to be expected, considering that Traneberg Bridge is more centrally located, with more alternative routes closer at hand.

Table 5.10. Comparison of average travel time, trip length and travel speed in Scenarios 3, 6 and 7 on a daily basis (changes in relation to Base Scenario)

Information	Scenario 3: Essinge Route 3		Scenario 6: Traneberg Bridge		Scenario 7: Danvikstull Bridge	
Travel time						
Mean change [min/trip]	+3.7	+15%	+0.5	+2%	+1.2	+5%
Max. change [min/trip]	+30.2	+55%	+7.7	+19%	+30.5	+127%
Trip length						
Mean change [km/trip]	0.0	–	+0.2	+1%	+0.2	+2%
Max. change [km/trip]	+1.0	+3%	+3.3	+11%	+7.8	+27%
Travel speed						
Mean change [km/h]	-4.5	-13%	-0.4	-1%	-1.1	-3%
Max. change [km/h]	-17.9	-37%	-3.0	-7%	-11.7	-25%

5.3.3 Bad Weather

In two of the scenarios, snow is assumed to cause capacity reductions in the road network. In Scenario 8 the whole network is subjected to a 15% reduction in free-flow speed, while the main road network (= superior network + primary and secondary network, including European highways, main arterial roads, county roads in urban or rural areas; local municipal links and city centre are excluded) is left unaffected in Scenario 9. The presentation (Table 5.11) is based on the morning rush hour, simulating the effects of a snowfall very early in the morning, i.e. just before and then a short time after snow clearing has begun.

As expected, trip lengths do not change considerably, while travel time increases when free flow speed decreases. The difference between the whole and only part of the road network being affected is quite large. In Scenario 8 the 15% decrease in free flow speed causes travel time increases of 15–20% (10–15 min) in about 40% of the travel relations. For about one third of the OD-relations the relative increase is even greater (maximum around 25%). In Scenario 9, however, the increase in travel time is only 2–4 minutes between a majority of the zones and the maximum relative increase (as opposed to the maximum absolute increase of 6% presented in Table 5.11) is about 9%. The prioritised network amounts to around 1/3 of the total model network in length, but generates almost 3/4 of the vehicle kilometres in the system. The effect is (as usual for this time of day) greatest in south-north relations. Also, travellers to/passing

e.g. the city centre are worse off, which is natural considering the roads included in the prioritised network.

Table 5.11. Comparison of average travel time, trip length and travel speed in Scenarios 8 and 9 during the morning rush hour (changes in relation to Base Scenario)

Information	Scenario 8: General reduction in capacity		Scenario 9: Selective reduction in capacity	
Travel time				
Mean change [min/trip]	+4.6	+17%	+1.2	+4%
Max. change [min/trip]	+14.9	+23%	+3.8	+6%
Trip length				
Mean change [km/trip]	0.0	–	+0.1	+1%
Max. change [km/trip]	-0.6	-1%	+1.0	+2%
Travel speed				
Mean change [km/h]	-5.2	-14%	-1.4	-4%
Max. change [km/h]	-8.9	-19%	-2.3	-5%

When considering the changes in speed in terms of percentages, it is found that a 15% reduction in free flow speed on all links (Scenario 8) gives rise to reductions in travel speed below that value just about as often as above it. When only part of the network is affected (Scenario 9), 95% of travel relations experience travel speed losses of only 6% or less. The brief conclusion drawn from this is that reductions in capacity due to weather can result in considerable delays. However, it is possible to reduce these effects to quite an extent by e.g. a well-planned snow clearing prioritisation.

5.3.4 Varying Link Capacity

The Essinge Route is Stockholm's most important arterial road. At the time of the study it had three lanes in each direction.[4] Closing one out of three lanes in the northbound direction (Scenario 1) reduces the number of vehicles passing the bridge at Gröndal in that direction during the morning rush hour from 5100 to 3850 vehicles, which in turn results in an increase in traffic volume from 1700 to 1925 vehicles per remaining lane. By comparing the results of closing one lane (Scenario 1) and all three lanes (Scenario 2), we can study the effect of different degrees of capacity reductions (Table 5.12).

[4] Two additional lanes, one in each direction, have been added in early 2002.

Table 5.12. Comparison of average travel time, trip length and travel speed in Scenarios 1 and 2 during the morning rush hour (changes in relation to Base Scenario)

Information	Scenario 1: Essinge Route 1		Scenario 2: Essinge Route 2	
Travel time				
Mean change [min/trip]	+0.1	+1%	+4.1	+15%
Max. change [min/trip]	+1.7	+4%	+39.9	+91%
Trip length				
Mean change [km/trip]	0.0	–	+0.1	+1%
Max. change [km/trip]	+0.2	+1%	+1.5	+5%
Travel speed				
Mean change [km/h]	-0.2	0%	-4.4	-12%
Max. change [km/h]	-1.4	-4%	-19.0	-41%

Table 5.12 clearly indicates that the effect of closing all three lanes is much worse than three times the effect of closing one lane. The resulting effects of Scenario 1 are however questionable. The road in question is very heavily trafficked and deemed to operate close to its capacity limit, and hence extremely sensitive to disturbances. Therefore one would expect larger effects in Scenario 1 than those indicated in the table.

The resulting volume of 1925 vehicles per hour and lane in Scenario 1 is just under the reference capacity of 2000. It is hence on the polynomial side of the vd-function and the problem discussed earlier (see Section 5.1.2) should not be an issue in this case. However, the results will be strongly dependent on how well the vd-functions are calibrated. It is obvious that the calculated effects may be smaller than expected simply because the vd-functions do not describe traffic conditions well enough. In other words, the curves may be too flat near the reference capacity, which leads to an underestimation of the effects resulting from an increase in traffic load. Hence there is probably a need for revising applied vd-functions.

5.3.5 Travel Demand Variation

The OD-matrices used to describe travel demand are determined by combining various sources of information (including traffic counts, travel surveys etc.) and are supposed to represent the average traffic volume during the chosen time period. In reality, traffic fluctuates more or less randomly around these levels, and the effects of these variations on e.g. travel times cannot be expected to be linear. To study this, the last three scenarios involve all entries of the OD-matrix being altered by -8%, +8% and +16%, respectively (Table 5.13).

Table 5.13. Comparison of average travel time, trip length and travel speed in Scenarios 10, 11 and 12 on a daily basis (changes in relation to Base Scenario)

Information	Scenario 10: Car traffic 1		Scenario 11: Car traffic 2		Scenario 12: Car traffic 3	
Travel time						
Mean change [min/trip]	-0.5	-2%	+0.5	+2%	+1.1	+4%
Max. change [min/trip]	-3.6	-6%	+2.3	+4%	+7.6	+14%
Trip length						
Mean change [km/trip]	0.0	–	0.0	–	0.0	–
Max. change [km/trip]	-0.2	-1%	+0.4	+1%	+0.6	+1%
Travel speed						
Mean change [km/h]	+0.6	+2%	-0.7	-2%	-1.5	-4%
Max. change [km/h]	+3.4	+7%	-1.8	-4%	-5.1	-12%

Regarding travel time, the relationship seems near to linear between decreasing and increasing the OD-matrix by the same factor (8%). There is nevertheless a faint tendency, in 2/3 of the travel relations, that the gains at lower traffic levels are a little less than the losses from a corresponding increase in traffic load. Hence, instead of cancelling out compared to the Base Scenario, the average of Scenario 10 and 11 tips over to "favour" our hypotheses of non-linear effects from variations in travel demand. This is even clearer when comparing Scenario 11 to Scenario 12. In the latter case, the travel time is more that twice as high as in the former case, in almost 90% of the travel relations. This non-linear effect is not so clearly observed for trip length or travel speed.

Generally speaking, the closer to its capacity limit a link operates, the greater is the non-linearity. In fact, only a few links are close to their reference capacities, which could explain the relatively speaking small average effects at the regional level. This can, on the other hand, also be caused by the way the vd-functions are constructed around reference capacities, as mentioned in previous sections. There is hence a risk that the link travel time effects of load increases will be underestimated.

5.4 Application of an Alternative Model

5.4.1 General

As discussed in Section 5.1.2, the EMME/2 solution algorithm is based on link flows only, and links downstream from overloaded links are also experiencing the extra queuing time. Also, the construction of the vd-functions beyond reference capacity means that all traffic being let through is just a question of time, and hence no actual capacity limit is imposed.

The Disaggregate Simplicial Decomposition – Implicit Route Storage model (DSD-IRS) (Larsson and Patriksson 1992; Tatineni et al. 1998) also solves the traffic assignment problem according to the user equilibrium principle, but is not based on link flows but on route flows. Therefore it can keep track of all routes used in every unique OD-relation. In this model system it is also possible to take capacity restrictions into account, by exchanging the linear part of the vd-functions for a deterministic queuing time (its length depending on the time period for which the oversaturated conditions are assumed to last), to be added to travel time on identified "bottlenecks" only. In this way, there is no double counting through undue "punishing" of traffic on downstream links. On these, the traffic flow is set to reference capacity – which in this case actually is a capacity limit – and travel time is calculated accordingly. Effects upstream from the bottlenecks are not modelled in any other way, than that the extra travel time imposed on the critical link will cause alterations in route choices elsewhere in the network. It should also be noted that DSD-IRS is subject to the same shortcomings as EMME/2 in its handling of trip timing and peak spreading.

A small complementary study using DSD-IRS was performed, to see how big a difference this improvement of the travel time calculations would make. Included Scenarios are 1, 2, 4, and 5, as well as the Base Scenario for the morning rush hour.[5] Results for the Base Scenario by this method are practically the same as for EMME/2. The outcome in the other scenarios is all but obvious, since the alteration may lead to downstream links being less congested, making average effects smaller. On the other hand, we cannot say if the overall redistribution of traffic will affect the system in the other direction. The results are summarised in the following, while detailed tables can be found in Berdica (2000).

5.4.2 Saltsjö-Mälaren Passage

Compared to Table 5.9, the changes in average travel time in Table 5.14 show an overall increase at the regional level. The maximum values are also considerably higher. The more detailed tables in Berdica (2000) display the same patterns, with the most affected travel relations being from the zones south of, to the zones north of, the Saltsjö-Mälaren passage. For Scenario 2, however, the increase in average travel time in relations from the South inner city to the zones north of the water strait is about 14% less according to the DSD-IRS model. Higher values from EMME/2 may well

[5] Stopping criteria and degrees of convergence are similar to the EMME/2 runs.

be the result of the "double-counting" feature described above. The differences in trip distance change between the models are negligible, while travel speed reductions are generally greater for the DSD-IRS model for all three scenarios – for all but the previously mentioned travel relations in Scenario 2, which is to be expected.

Table 5.14. Comparison of average travel time, trip length and travel speed in Scenarios 2, 4 and 5 during the morning rush hour (changes in relation to Base Scenario) using DSD-IRS

Information	Scenario 2: Essinge Route 2		Scenario 4: Central Bridge		Scenario 5: Western Bridge	
Travel time						
Mean change [min/trip]	+4.6	+16%	+2.4	+9%	+0.7	+3%
Max. change [min/trip]	+44.7	+83%	+24.7	+55%	+9.3	+21%
Trip length						
Mean change [km/trip]	+0.1	0%	+0.1	+1%	0.0	–
Max. change [km/trip]	+1.6	+5%	+2.1	+8%	+0.7	+5%
Travel speed						
Mean change [km/h]	-5.0	-14%	-2.6	-7%	-0.9	-3%
Max. change [km/h]	-20.3	-44%	-10.8	-23%	-6.6	-18%

5.4.3 Varying Link Capacity

Comparing the DSD-IRS runs of Scenarios 1 and 2 to each other shows, just like the EMME/2 results, that a complete closure in the northbound direction is not proportional to closing only one out of three lanes. Much more interesting, however, is the difference in Scenario 1 using the two models. Although the average changes in Table 5.15 do not appear to differ very much, a closer look at a more detailed level shows that DSD-IRS yields travel time increases of 3–4 times those of EMME/2, in affected travel relations. The corresponding differences for Scenario 2 are not so great. Hence it seems that DSD-IRS takes better account of partial closures of a link.

5.5 Conclusions

It can be discussed to what extent it is possible to use the link flow based user equilibrium model EMME/2 to study vulnerability issues. Since bottleneck effects are not taken into consideration, modelling partial capacity reductions (e.g. closing one of three lanes) seems to give unrealistically small effects. This is supported by the complementary study using the route flow based DSD-IRS model. Also, using EMME/2 to study effects of

drastic capacity reductions may be to push the model beyond the boundaries of its validity, since queuing times when demand exceeds reference capacity are difficult to represent in a realistic way. Then again, effects being intuitively too small could be a question of vd-function calibration in general – they may be too flat near the reference capacity, which if so leads to an underestimation of travel time increases.

Table 5.15. Comparison of average travel time, trip length and travel speed in Scenario 1 (changes in relation to Base Scenario) using EMME/2 and DSD-IRS, respectively

Information	Scenario 1 using EMME/2		Scenario 1 using DSD-IRS	
Travel time				
Mean change [min/trip]	+0.1	+1%	+0.4	+1%
Max. change [min/trip]	+1.7	+4%	+4.5	+10%
Trip length				
Mean change [km/trip]	0.0	–	0.0	–
Max. change [km/trip]	+0.2	+1%	+0.1	+1%
Travel speed				
Mean change [km/h]	-0.2	0%	-0.5	-1%
Max. change [km/h]	-1.4	-4%	-3.6	-9%

It is, however, important again to remember the conditions and assumptions that lie behind the present approach to the problem. The user equilibrium assignment means that all travellers have chosen their minimum time route, with respect to a system status that is assumed to be known to them beforehand. Therefore one cannot expect the model results to display the same drastic travel time increases that result from the immense queues caused by unforeseen incidents in real life traffic. Modelling the effects of such sudden interruptions calls for micro simulation models. So far, this has not been a realistic approach for large-scale regional applications to a city like Stockholm. This will change in the future, since micro simulation models are now being implemented on a regional scale. Moreover, to study the most interesting effects of sudden incidents on the critical bridges in the central area, it may not be necessary to include the whole of the Stockholm Region in the study area. Studies similar to the present one, using a micro simulation approach, have been performed elsewhere (Nicholson et al. 2001; Berdica et al. 2002). The road system was then much smaller: an area of roughly 16 km^2 compared to 6,500 km^2 in Stockholm, and accommodating only a tenth of the number of trips during the morning rush hour. It should also be remembered that the present method gives the traveller only one option for resolving their "trip problem", and that is the choice of an alternative route. In reality, some may change their mode of transport, choose to travel at a different time of day or maybe even change their des-

tination (e.g. shop at a different super market), which is why the results cannot be transferred to a real life situation on a one to one scale. As was mentioned in previous sections, the excessive link flows, which quite obviously cannot pass during the time period modelled, are to be interpreted as the demand for passage on that particular link.

Nevertheless, keeping these limitations and reservations in mind, there are some points to be made from the impact analysis of the various scenarios:

- The bridge at Gröndal (Essinge Route), Western Bridge and Central Bridge can be regarded as main alternative routes for each other in the case of an incident.
- For a closure in the northbound direction during the morning rush hour, the Essinge Route is the most critical one as far as consequences are concerned, resulting in 20–40 min extra per trip across the Saltsjö-Mälaren passage.
- Closing Danvikstull Bridge results in daily travel time increases of well over 20 minutes for most journeys to/from the Eastern Region. This is quite different from the effect of just over one minute per trip at the regional level.
- Snow/sleet in the whole road network (=15% decrease in free flow speed) during the morning rush hour causes travel time increases of 10–15 minutes in about 40% of the travel relations. Keeping a priority network "clear" reduces this to only 2–4 minutes extra between a majority of the zones.
- There is a faint tendency of non-linear effects from variations in traffic demand. This shows most clearly when increasing traffic by 16%, which results in more than double the extra travel time caused by increasing traffic by 8%.
- In virtually all scenarios there are both winners and losers, which in turn results in the average effects at the regional level sometimes seeming insignificant, in terms of absolute figures. On the other hand, they can be revealed to be quite large when related to normal conditions.
- Calculating effects in monetary terms (35 SEK/h, 1.3 SEK/km; 1999) shows extra costs ranging from 200 000 to about 5 million SEK per day, depending on which scenario is considered. The corresponding costs, were the situations modelled as "unforeseen", are probably considerably higher.
- According to these calculations, keeping the main road network "clear" from snow and sleet leads to an estimated daily travel cost saving of about 3 million SEK.

Even though the numbers stated above are subject to uncertainty, they clearly indicate the Stockholm road transport system's weaknesses, and give a fair idea as to the relative size of the expected consequences. Translating this into monetary terms can then give values of a wide range, depending on how one values the effects, for which there is no consensus. Setting aside the specifics, this model-based case study has shown that the central water strait – and the consequent dependency on a limited number of bridges – is a crucial feature from a vulnerability point of view. Also, there is at present little, if any, extra capacity to handle rerouting in case of incidents, let alone to swallow a more permanent increase in traffic volume. This puts high demands on the operations management and maintenance of the road (and public transport) infrastructure by e.g. detecting and attending to occurring disturbances within a minimum time span. These reactive measures must be combined with the proactive approach of identifying critical links that are to be prioritised for enhancement. Also, the undertaking of any measure should be planned carefully to minimise disturbances from the roadwork itself.

A more general conclusion is that the transport models and methods of today can be used to conduct vulnerability studies, although they are not yet ideally adjusted to suit this specific purpose. From the vulnerability perspective it is crucial how well the model reproduces effects around reference capacities, so in the short term the currently used vd-functions should be revised. In the longer perspective, there is a need for developing better tools in general, probably in the form of micro simulation approaches. When using network equilibrium models, it is important that the assignment problem is solved accurately. Considering in this case Stockholm's growing population and the continuous increase in traffic, the importance and usefulness of further engagement in this area is evident.

Acknowledgments

The authors are grateful for helpful comments from the editors and from two anonymous referees. The research was given financial support from the Swedish Agency for Innovation Systems, the Swedish Agency for Civil Emergency Planning and the Swedish Road Administration.

References

Ball MO, Golden BL, Vohra RV (1989) Finding the most vital arcs in a network. Operations Research Letters 8: 73–76

Berdica K (2000) Analysing vulnerability in the road transportation system: putting theory into practice using Sweden as an example. TRITA-IP FR 00-76, Department of Infrastructure, Royal Institute of Technology, Stockholm

Berdica K (2002) An introduction to road vulnerability: what has been done, is done and should be done. Transport Policy 9: 117–127

Berdica K, Andjic Z, Nicholson A (2002) Simulating road traffic interruptions: does it matter what model we use? In: Bell MGH, Iida Y (eds) The network reliability of transport. Elsevier, Amsterdam, pp 353–368

INRO (1998) EMME/2 User's manual release 9. INRO Consultants Inc., Montreal

Jenelius E, Petersen T, Mattsson L-G (2006) Importance and exposure in road network vulnerability analysis. Transportation Research A 40: 537–560

Lam, WHK (1999) Special issue: network reliability and transport modelling, editorial. Journal of Advanced Transportation 33: 121-123

Larsson T, Patriksson M (1992) Simplicial decomposition with disaggregated representation for the traffic assignment problem. Transportation Science 26: 4–17

Nicholson A, Berdica K, Andjic Z (2001) Comparing traffic models: two case studies. Transport Engineering in Australia 7: 65–76

Sheffi Y (1985) Urban transportation networks. Prentice-Hall, Englewood Cliffs, New Jersey

Tatineni M, Edwards H, Boyce D (1998) Comparison of disaggregate simplicial decomposition and Frank-Wolfe algorithms for user-optimal route choice. Transportation Research Record 1617: 157–162

Taylor MAP, D'Este GM (2004) Critical infrastructure and transport network vulnerability: developing a method for diagnosis and assessment. In: Nicholson A, Dantas A (eds) Proceedings of the second international symposium on transportation network reliability (INSTR). Christchurch, New Zealand, pp 96–102

Transek (1999) Car-users' choices in road traffic: a study based on focus group discussions and Stated Preference methodology (in Swedish). Transek AB, Solna

Wallman C-G (1996) Effect calculations for the "ready reckoner": speed reductions and fuel consumption during different road weather conditions (in Swedish). VTI-notat 71-1996, Swedish National Road and Transport Research Institute, Linköping

Wollmer R (1964) Removing arcs from a network. Operations Research 12: 934–940

6 Survivability of Commercial Backbones with Peering: A Case Study of Korean Networks

Morton E. O'Kelly and Hyun Kim

Department of Geography, The Ohio State University, Columbus, OH, 43210, USA; Emails: okelly.1@osu.edu; kim.1567@osu.edu

6.1 Introduction

The Internet is a key technology for the information age. It has evolved as a fundamental communication mechanism across a wide range of sectors (academic, personal, business, government, etc.) and spans the globe. This worldwide communication system is operated by multiple network providers. Internet Backbone Providers (IBPs) manage their own long-haul transmission networks with high bandwidths to transit Internet traffic and a number of Internet Service Providers (ISPs) mainly provide customers with access by connecting them to IBPs (Malecki 2002; Dodd 2002).

Internet traffic needs to be linked across networks for users to gain access to the services in other geographical regions. *Internet hubs* such as Internet eXchange points (IXs) and Network Access Points (NAPs) refer to geographical common nodes where multiple providers can exchange their traffic. *Peering* is defined as a way to interconnect multiple network providers in these hubs. These access points are generally located in highly accessible city nodes for the efficiency of traffic exchange among peering members.

Maintaining the Internet at a high level of reliability is a major challenge. Specifically, the protection and placement of hubs has been identified as a critical defensive strategy since networks are so dependent upon their operation. Failure of hubs from intended or unintended disruptions can cause disastrous impacts on telecommunication ability (Grubesic et al. 2003; NSTAC 2003; NCA 2003).

Korea experienced a major incident in January 2003 called the '*1.25 Worm Incident*' associated with concentrated worm virus attacks. The critical failure of the core servers of *Korea Telecom* (KT), the biggest Internet backbone provider, caused the temporary but significant malfunction of various Internet services at the national level. Several commercial backbone providers and local ISPs relying on the KT network suffered from cascading degradation of their Internet services. Damage occurred in the critical Internet hub located in Seoul where over 70 ISPs are highly interconnected via peering arrangements (NCA 2004a). This incident stresses the vulnerability of the Internet to unexpected damage at critical nodes, and in turn highlights the importance of examining Internet resilience from a geographical perspective.[1]

The main purpose in this chapter is to explore the vulnerability of the Internet for city nodes of Korea, taking into account peering arrangements in select hubs. Particularly, *network reliability* is used as a measurement to examine vulnerabilities of the Internet. A novel concern of this chapter is to examine the performance of individual ISPs, and the resiliency of Korean networks for the hypothetical intended hub (nodal) attacks on major ISPs' access points, as well as *Internet eXchanges* which play a role connecting foreign countries.

This chapter also explores how patterns of vulnerability differ between Korea and the United States. Different geographies [spatial structure] of the Internet could lead to different spatial patterns and implications (Huh and Kim 2003). The chapter is divided into five sections. Within section 6.2, the issues in examining the vulnerability of the Internet and relevant literature are explored. Section 6.3 explains the characteristics of Internet hubs and the measurements applied in our model. Emphasis is placed on the idea of utilizing a *reliability envelope* as an analytical tool. Pertinent background information describing the geography of the Internet in Korea is provided in section 6.4. The results gathered from empirical analyses are also presented in Section 6.4, followed by concluding remarks in the final section.

[1] Literature related to the vulnerability of the Internet generally focuses on the failure of lower layers of the Internet protocols such as the disruption of physical network components. Even though higher level layers such as application and presentation may not be directly impacted by these failures, these losses could influence the reliability of entire systems. For example, the higher level of layers in the OSI model could suffer from traffic congestion due to the collapse of the physical layer (Moore et al. 2002; Broadband Week 2001).

6.2 Vulnerability of the Internet

Network vulnerability receives attention from many fields, particularly those dealing with network design since it is required to embed vulnerability measurements in designs for reliable or survivable networks. Since early network designs were built with redundancy in mind (Baran 1964), various approaches have been proposed to find critical network components, as well as assess the system's vulnerability (Houck et al. 2003; Ellison 1999; Ball et al. 1989; Baybars and Edahl 1988). The degree of vulnerability of infrastructure and associated measurements were briefly discussed by Shake et al. (1999). The possible consequences of cascading failures of critical infrastructures have been discussed in recent literature (White House 2003; Little 2002; Carreras et al. 2002).

In recent years, geographers have begun to pay close attention to vulnerabilities of the Internet. Previous work has used such measurements as connectivity, graph theoretical indices, and nodal accessibility (O'Kelly and Grubesic 2002; Malecki and Gorman 2001; Moss and Townsend 2000). An effort to identify vital nodes in U.S. critical infrastructure was made by Gorman et al. (2004) through the creation of a database of national data carriers. More recently, the potential impacts from failures of vital nodes in geographically linked network were explored by applying a spatial optimization model (Grubesic and Murray 2006). In the case of non-western countries, the urban hierarchy, as well as the spatial structures of the Internet backbone in Korea, is explored through the application of network analyses (Huh and Kim 2003).

The main concern of these studies was to assess the *potential* availabilities at both individual nodes and in commercial Internets. This research has been conducted under the common assumption that all network components of the Internet should operate normally without any failures. Recent work (Grubesic et al. 2003) has addressed survivability as another approach to examine vulnerability of the Internet in the U.S. They show how the survivability of the commercial Internet in the U.S. would change if some selected nodes were disrupted. Specifically, they defined survivability as the smallest amount of damage to cause disconnection of a network. These analyses were conducted under an *all-or-nothing* assumption, so that the realistic characteristics of telecommunication networks, such as multiple interconnections among backbone providers, have been largely ignored or reflected in a simplified manner.

In more recent studies, the concept of *reliability* has been applied in examining the vulnerability of U.S. cities for select commercial Internet backbones (O'Kelly et al. 2006). They suggest that the vulnerability meas-

ure should be refined by taking into account *peering arrangements* at Internet hubs. As disconnections increase, network reliabilities will be degraded *probabilistically*. In particular, the idea of a *reliability envelope* has been devised to reflect important network properties in terms of reliability characteristics. Their work is more focused on the reliability of domestic Internets based on selected hubs in U.S.

6.3 Reliability, Peering Arrangement in Internet Hubs and Reliability Envelope

6.3.1 Network Reliability

The concept of reliability has been addressed in various fields such as system engineering and telecommunication network design in terms of connectivity and traffic management (see Baran 1964; Colbourn 1987). Network reliability is generally defined as the ability of a network to carry out a desired network operation despite the failure of network components, such as nodes or linkages. In contrast to the other deterministic measures of vulnerability, the performance of a network providing successful communications is presented in the form of probabilities. Therefore, it is assumed that any network component operates according to certain known probabilities. It is necessary to explain how to compute '*Origin - Destination*' reliability, the probability of operation based on the paths connecting two nodes, since our focus is on assessing the reliability of specific pairs of city nodes on the Internet. The most fundamental method of making such measurements is to sum up probabilities of all disjoint events between two nodes after completing all possible state enumerations. The mathematical expression is as follows (Shier 1991; O'Kelly et al. 2006):

$$R_{OD}(G,p) = \sum_{i=1}^{n} P\{\delta_i\}$$

$$P\{\delta_i\} = \prod_{j=1}^{m} P\{e_{ij}\}$$

Where

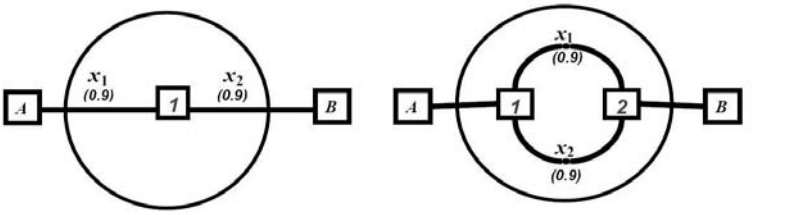
R_{OD} : the reliability for two nodes, origin O and destination D

G : a graph of a network with parameter p (known probability) for edges

$P\{\delta_i\}$: the probability of the disjoint event δ_i

e_{ij} : the edges j constituting δ_i of states

Two fundamental computational cases of O-D reliability are illustrated in Figure 6.1. Using a given parameter of '$p = 0.9$' as an operational probability for each linkage, the reliability of a serial arrangement is calculated by multiplying probabilities of each linkage consisting of the disjoint event $\delta_1 \{x_1, x_2\}$. In the case of a parallel arrangement, the inclusion-exclusion method is commonly employed to exclude double counting sets. In a given parallel system (Figure 6.1(b)), the probability of intersection event $E_1 \cap E_2 = \{x_1, x_2\}$ is deducted from the unions of two successful events E_1 and E_2 to make disjoint events. However, this technique is known to be impractical in the case of a large network.[2]



(a) Serial arrangement

$E_1 = \{x_1, x_2\}$

$R_{AB} = \Pr(E_1)$

$= 0.9*0.9 = 0.81$

(b) Parallel arrangement

$E_1 = \{x_1\}, E_2 = \{x_2\}$

$R_{AB} = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$

$= 0.9 + 0.9 - (0.9*0.9) = 0.99$

Fig. 6.1. Reliability computation for simple arrangements (*Source*: O'Kelly et al. 2006)

As an alternative, a Boolean algebra method has been employed in this study. The Boolean method computes the exact reliability by enumerating events using by a disjoint product technique (Ball et al. 1995; Shier 1991; Yoo and Deo 1988; Agrawal and Barlow 1984; Fratta and Montanari 1973). The main idea of this method is to work forward from an initial successful event and to search another disjoint event by utilizing its complement and other unused viable paths, based on Boolean logic, until no more disjoint events are found. For example, assuming that there are three successful events, E_1, E_2 and E_3, then disjoint events δ_i by Boolean logic are assigned as $\delta_1 = E_1$, $\delta_2 = \overline{E_1} \cap E_2$, and $\delta_3 = \overline{E_1} \cap \overline{E_2} \cap E_3$. Since

[2] Basically, since the computation of reliability at least follows the complexity of $O(2^n)$ [n is the number of events], various techniques for reducing this complexity have been proposed based on the Inclusion-exclusion method. However, this approach is only effective in a few specific cases.

these are mutually disjoint, the summation of probabilities of these events represents the reliability between two nodes. Though this procedure is different from a basic reliability computation, it is known to be more efficient than other exact methods. Figure 6.2 illustrates how this algorithm finds the disjoint events in the case of a parallel arrangement as compared to the opposing method shown in Figure 6.1(b).

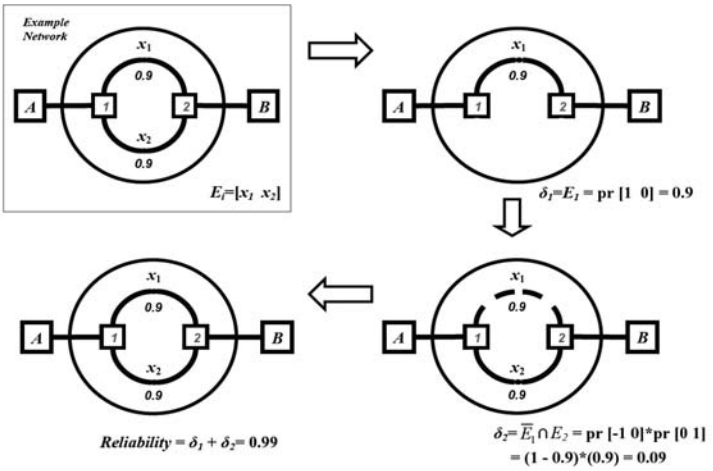


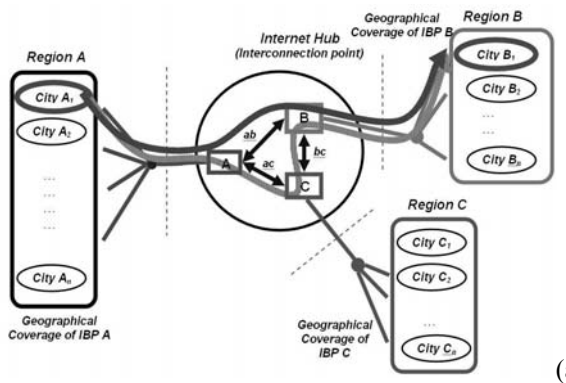
Fig. 6.2. Reliability algorithm using Boolean algebra in a parallel arrangement (*Source*: O'Kelly et al. 2006)

6.3.2 Peering Arrangements in a Hub

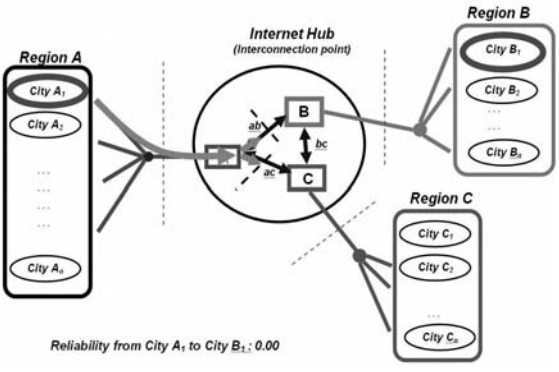
Peering occurs between at least two ISPs as a cooperative relationship. These arrangements are generally formed for such particular city nodes as Chicago (Network Access Point), Dallas (Metropolitan Area Exchange) and Seoul (Internet eXchange). Employing these mutual interconnections, even small ISPs can have geographically expanded service coverage and are more reliable *via* redundant paths in the event of operational problems.[3]

[3] More precisely, there are two types of arrangements. *Peering* is generally a mutual arrangement between similar sizes of ISPs for gaining economic advantages rather than competitive advantages by directly exchanging their traffic. In contrast, *transit* refers to the service being provided to the small ISPs by large ISPs to deliver traffic from the smaller ISPs to other Internets (Huston 1999). Most small and local ISPs in Korea usually utilize the transit service via the major ISPs backbones. Even though they should pay transit fees according to contract, however, they can extend their service area as well as enjoy the economic benefits

Figure 6.3 illustrates how peering arrangements work among participant Internet backbone providers. Suppose that there are three Internet Backbone Providers, A, B and C. Each takes charge of a geographically separated service area. A single exchange point is established as an Internet hub where these IBPs are participated for the purpose of peering. As shown in Figure 6.3(a), mutual interconnections among participants enable them to exchange traffic within the hub. For instance, the traffic originated from city A_1 can be transmitted into the targeted city B_1 through the Internet hub, since peering among these participants is operational. The diagram shows how the users subscribing to IBP A can access the information source located in city B_1 which is managed by IBP B by virtue of a peering arrangement. The important characteristics from a given peering arrangement are stressed in terms of network reliability.



(a)



(b)

Fig. 6.3. Function (a) and malfunction (b) of the peering arrangement in a hub

of the better quality and reliable service. KT and Dacom are known as the biggest IBPs among major ISPs providing both peering and transit services (NCA 2004).

Firstly, the Internet can be made more resilient due to the created sub-network within the hub. In Figure 6.3(b) for instance, successful communication between cities A_1 and B_1 (Figure 6.3(a)) is possible if at least one of the inter-linkages works. Secondly, the type of peering implemented can also influence network reliability. If all participants in a hub are completely interconnected to each other (public peering), the network would be more reliable since a small number of disconnections might not significant impact on the traffic exchanges between participants. In the case of private peering, Internet traffic is allowed to be carried only through dedicated interconnections for ISPs, according to their bi-lateral contracts.

The main reason for private peering is to deliver a large amount of traffic without congestion between particular ISPs. In reality, peering relationships are formed in a more complicated fashion, as exemplified in Figure 6.4. For instance, the peering relationship of *Dacom* (BORANET), the second largest ISP in Korea, is formed by the link of two Internet eXchanges (*KIX* and *KTIX*) which allows public peering, and also facilitates private peering with over 20 ISPs. An ISP with private peering can communicate with peering members during the operational failure of public hubs. Since the Internet can retain its operational functionality by virtue of peering arrangements until the hub reaches a critical threshold, reliabilities for pairs of cities on the Internet should experience degradation of service caused by malfunctions of linkages in a hub, rather than an abrupt, complete loss of telecommunication ability. The idea of a *reliability envelope* is appropriate to reflect these realities.

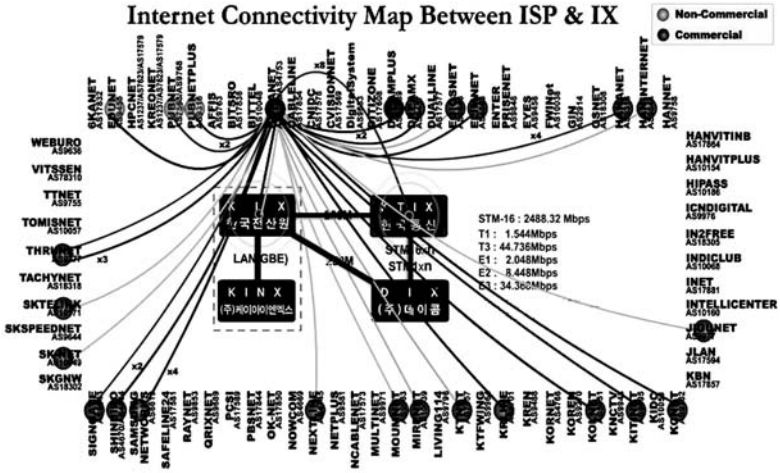


Fig. 6.4. A case of peering relationship (BORANET)
(*Source*: http://isis.nic.or.kr/sub03/sub03_index.html)

6.3.3 Reliability Envelope

The all-or-nothing principle assumed in prior research has been applied in order to measure network survivability by showing the potential for network disruption resulting from complete damage on network components, such as particular nodes and linkages. This simplified assumption is relaxed here by considering all possible disconnections in peering arrangements enumerated by combinatoric rules. Particular peering arrangements in selected hubs are represented as a set of sub-networks. As the number of damaged linkages of the total interconnections increases, the reliability of the sub-network is expected to fall below the tolerable level degrading the overall network reliability.

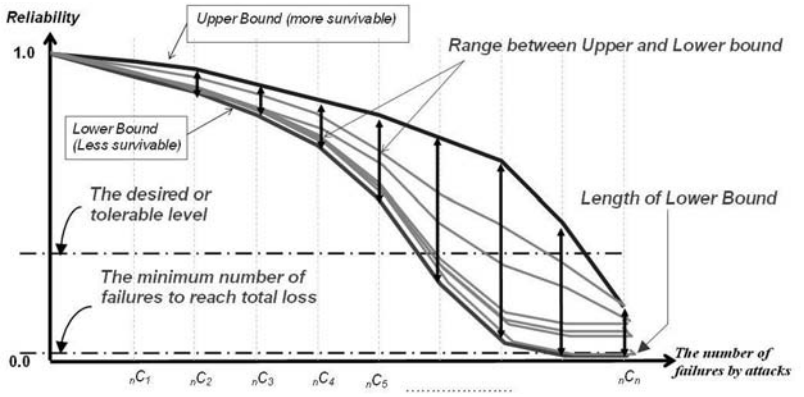


Fig. 6.5. Variations of reliabilities on the disruptions of hubs or interconnections in hubs

The reliability envelope in Figure 6.5 illustrates variations of reliabilities for a pair of city nodes on the Internet highlighting how a network can maintain its functionality in the failures.[4] When an attack impacts r out of n components at r stage ($r = 1$ to n), there is a range of consequences among the r chosen failures, from the relatively unimportant (being the best case scenario) to the highly damaging (being the worst case scenario).

[4] The conceptual difference between survivability and reliability relies on what circumstance is assumed for network operation. Reliability is associated with the ability of successful operation or adequate performance. In contrast, survivability deals with the ability to maintain its communicative capability focusing on whether a network can still be function in the face of failures (see Colbourn 1999; Soni et al. 1999; Gavish and Neuman 1992). In this sense, reliability envelope is a tool to analyze the degree of survivability of a network. Ideas similar to the reliability envelope are found in other literatures (see Doyle et al. 2005; Urban and T. Keitt 2001).

As the size of r relative to n increases, the chances that r failures produce minimal damage decline.[5] Based on this diagram, several analytical points can be surmised, as follows (based on O'Kelly et al. 2006):

[1] The ranges between upper and lower bounds at each of r stages show how vulnerable the sub-network is. If the network can keep the range of impact narrow even as disrupted components r increase, then by indication the network has a good resiliency against disruptions. The larger the range, however, the more susceptible the network is to failure. As illustrated in Figures 6.6(a) and (b), Figure 6.6(b) can be recognized as a more resilient hub than Figure 6.6(a) despite both reliability envelopes being limited by the same upper bound. This is because the range for Figure 6.6(b) is kept narrow for all stages indicating the network is more survivable from the possible attacks.

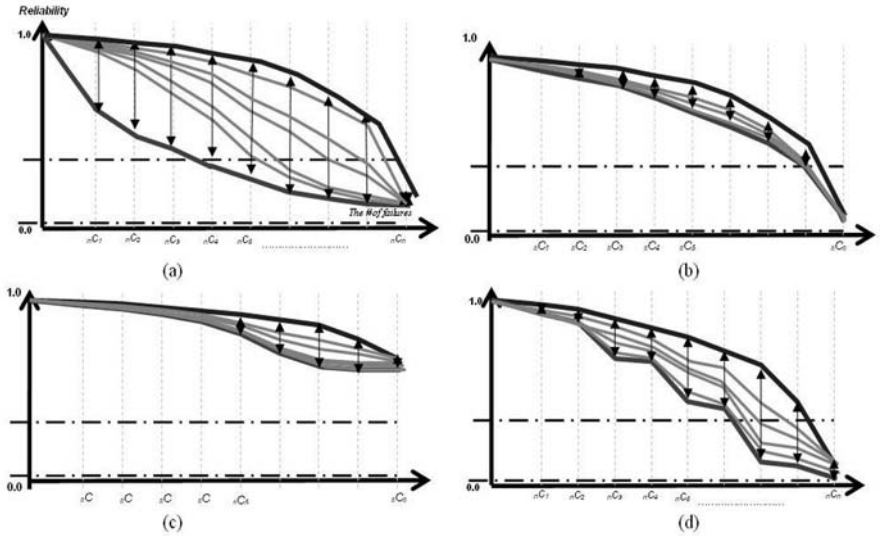


Fig. 6.6. Hypothetical shapes of reliability envelope

[2] The shape of the lower bound in a reliability envelope is discussed in terms of the survivability of a hub network. The length of the lower bound before reaching a complete loss of its functionality shows how long the Internet hub can withstand an increase of attacks. The longer the lower

[5] r can be interpreted as the ability of the attacker to damage the network components. If an attacker is highly intelligent to disrupt the r components causing significant damage on network performance, then it would be the best scenario in terms of an attacker. The reliability envelope therefore shows all potential consequences of failures from attackers.

bound before reaching the level of total loss, the more survivable the network. Compared to the previous case, the profile in Figure 6.6(c) would be more survivable during hub failures. The stepped shape in Figure 6.6(d) implies that particular combinations of disconnections might cause significant reliability degradation at select stages.

[3] The desired level of Internet reliability can be discussed in terms of network design. If multiple Internet hubs exist on the network, the most susceptible hubs causing significant degradation of reliability can be examined using a reliability envelope. A reliability envelope would allow for the exploration of such pertinent issues as determining the smallest number of interconnections needed to enhance the survivability of the network for attacks, or perhaps in determining the best peering arrangements needed to improve telecommunication ability.

6.4 Analysis

As a preliminary analysis, spatial pattern based on reliability in terms of city nodes under normal conditions is explored with a brief explanation of the data sets used in our analysis. The first focus of empirical analyses is placed on the reliability envelope for the lowest reliability city pair selected from preliminary analysis. Then, the performance of individual ISPs is compared based on their envelope profiles showing which ISP network would be relatively susceptible to the others. The final objective is to examine the geographic variations of reliabilities to be connected to a particular foreign country from potential disruptions.

6.4.1 Internet eXchanges, Peering, and Commercial Internets

Currently, four Internet eXchanges (IXs) exist to provide the exchange services for a number of ISPs in Korea. All four IX nodes are located in the Seoul metropolitan area, which is in direct contrast to the geographic layout of nodes in the U.S., where a number of access points such as NAPs are more geographically distributed. The regional traffic exchange point (R-IX) was recently established in Busan, the second largest city located on the south shore of Korea. It was integrated so as to improve Internet service in South Korea by mitigating the heavy traffic concentration in Seoul. The peering of this IX has not been considered for the purposes of our study due to insufficient information.

Figure 6.7 shows the connectivity among the four IXs. The role and characteristics of each IX is differentiated. KIX, the first IX, established in

1995, is dedicated to serving non-commercial ISPs, such as those of government, non-profit agencies and educational organizations, while most commercial ISPs receive their traffic exchange service from either KTIX (Korea Telecom Internet eXchange) or DIX (Dacom Internet eXchange). KINX allows its service only to consortium members consisting of small and mid-size ISPs. Of the 70 plus ISPs categorized/documented as of 2003, only a handful of ISPs are classified as major ISPs[6]. Our analysis considers the eight major commercial Internets, including Korea Telecom (KORNET), Dacom (BORANET), Onse (Shinbiro), Hanaro (Hananet), Thrunet, Enterprise ID, SK Telecom (SKSpeednet) and Dreamline (DreamX). Each ISP has its own backbone network, covering the entire country as its service area. Other mid and small ISPs generally take charge of a local coverage area by leasing part of these major ISPs.

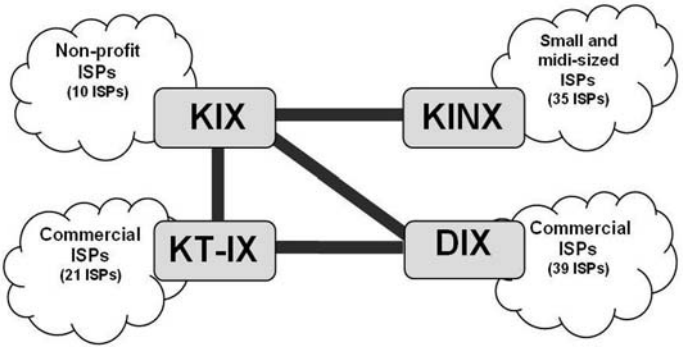


Fig. 6.7. Connectivity among *Internet eXchanges* in Korea (*Source*: NCA, 2003)

Figure 6.8 illustrates the topologies of the major ISPs (NCA 2003). Due to the availability of topologies and peering arrangements for these backbones, a fairly realistic analysis can be performed here. Since the focus of this study lies on commercial Internet sources, two IXs (KTIX and DIX) and the international connections from these ISPs to the U.S. are considered in our model.

[6] These major ISPs can be classified as IBPs; however, they are generally regarded as ISPs because they provide access service for both smaller ISPs and individual subscribers.

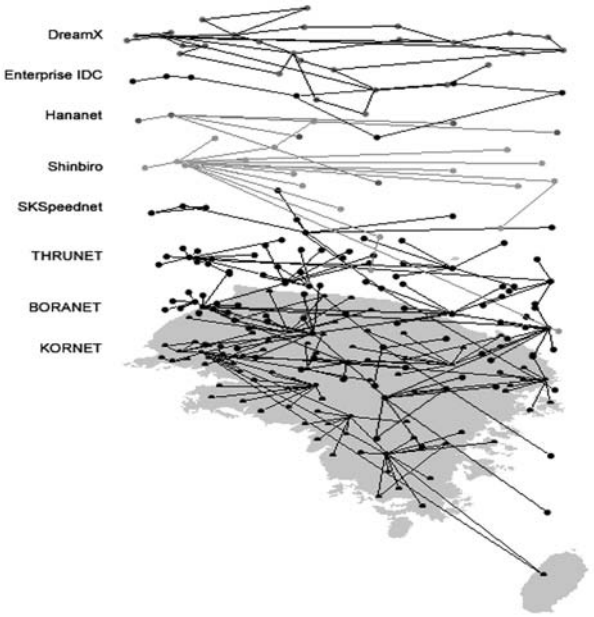


Fig. 6.8. Topologies of major ISPs in Korea

Figure 6.9 shows the peering relationships among the ISPs and both IXs. It also represents ISPs which act as a gateway to the U.S. As indicated in Figure 6.9, the actual peering relationships consist of a complicated network, made up of multiple assignments at public and private peering levels, and even the international connection level.

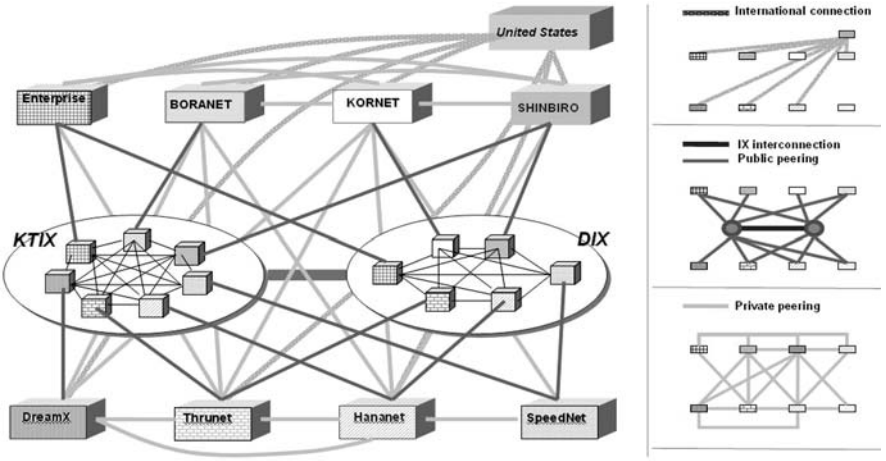


Fig. 6.9. Peering arrangements among two IXs and eight major ISPs

6.4.2 Accessibility and Reliability of Individual City Nodes

There exist a total of 91 city nodes among the eight major ISP networks. It is important to determine which cities are more accessible and/or more reliable; therefore, the nodal accessibility for 91 city nodes based on the overlaid network is measured by their total accessibility (Taaffe et al. 1996). The overall performance of each individual city node is then examined in terms of average reliability by computing the mean value of reliabilities from a particular city to the other cities, excluding that of the node being computed.

Table 6.1 Rankings of city nodes in Reliability and Nodal Accessibility

a) City node in top 10

	Nodal Accessibility		Average Reliability	
Rank	City node	T-Value	City node	Reliability
1	Seoul*	107,848	Seoul*	0.9487438
2	Daejeon*	98,803	Busan*	0.9487408
3	Anyang	78,566	Daegu*	0.9487361
4	Busan*	70,850	Incheon*	0.9487323
5	Daegu*	61,183	Gwangju*	0.9487274
6	Jeonju	57,163	Daejeon*	0.9487203
7	Gwangju*	54,819	Anyang	0.9487179
8	Wonju	46,373	Goyang	0.9487164
9	Seongnam	43,257	Cheongju	0.9487117
10	Suwon	41,198	Guri	0.9487107

(b) City nodes in bottom 10

	Nodal Accessibility		Average Reliability	
Rank	City node	T-Value	City node	Reliability
82	Cheorwon	5,872	Gimje	0.8547665
83	Yeoju	5,084	Buan	0.8547661
84	Hanam	5,084	Hanam	0.8547648
85	Paju	5,068	Taebaek	0.8547570
86	Yeongdong	4,065	Sokcho	0.8547569
87	Okcheon	4,065	Samcheok	0.8547569
88	Jeungpyeong	4,065	Hongcheon	0.8547567
89	Gimpo	2,623	Donghae	0.8547565
90	Yangpyeong	2,565	Cheorwon	0.8547564
91	Jinju	1,061	Jinju	0.8452156

* Metropolitan areas

As indicated by the results shown in Table 6.1, all five metropolitan areas examined in this study—excluding the Ulsan metropolitan area—are listed in the top 10 in both measure of total accessibility and reliability. The fact that Seoul is ranked number one in both accessibility and reliability stresses its critical role to the networks in Korea. It is interesting to point out that Incheon, a metropolitan area located adjacent to Seoul, is ranked 21st in terms of accessibility and 4th in terms of average reliability. Incheon's high ranking is due in part to the characteristics of the reliability measure. That is, the fewer steps between nodes, the higher the reliability generally computed, while the degree of accessibility is influenced more by the number of paths.

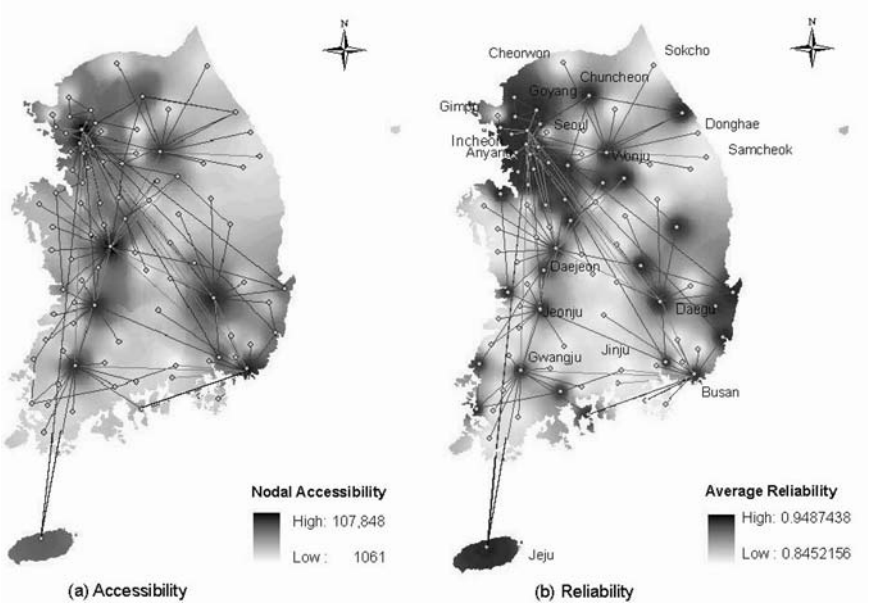


Fig. 6.10. Nodal accessibility and reliability potential map of Korea

The reliability potential map in Figure 6.10 (a) and (b) reveals a geographic pattern that such a classical measure as nodal accessibility might not catch, where either potentially vulnerable [bright] or more reliable [dark] cities are clustered. According to this map, more reliable clusters are formed around the five metropolitan areas (Seoul, Daejeon, Busan, Daegu, Gwangju and Incheon) and two regional centers, Jeonju and Wonju. The susceptible areas are located mainly in the mountainous regions that are found between the northeast coast to the south coast area. As indicated by both measures (reliability and accessibility), Korean Internet service providers rely heavily on Seoul as a critical node. The telecommunication

abilities of other Korean cities can therefore be critically degraded if malfunctions occur at peering arrangements in Seoul. It is important to simulate how and when reliabilities may be degraded with probable disruptions by examining reliability envelopes in terms of particular city pairs.

6.4.3 Performance of Individual ISPs

It is possible to examine reliability envelope profiles for all 4,004 possible city pairs, but this analysis will focus more on the select city pair of Seoul and Yangpyeong. This is one of the most vulnerable pair experiencing the significant degradation of reliability when a total malfunction on peering arrangements occurs in Seoul. Node attack could be a more plausible scenario rather than linkage attack since a node disruption more effectively threatens the network survivability even though the reliability envelope in terms of linkage failures in peering sub-network shows the reliability variation more precisely. In terms of computational time, the simulation of node failures is considered to be more practical since it can reduce the number of stages by removing inner nodes instead of complete linkage disconnections. Since there are ten hub nodes [8 ISPs : 2 IXs], only 10 stages should be considered when constructing a reliability envelope.[7]

Figure 6.11 shows the average reliability in terms of individual ISPs for the Seoul-Yangpyeong city pair. Under normal conditions (0 stage), all ISPs begin from a comparable level of reliability, indicating no difference of performance. However, Enterprise IDC, SKSpeednet and Hananet are expected to degrade considerably with the disruptions from the 5^{th} stage while other ISPs show patterns of gradual decline indicating relatively good performance and resiliency against disruptions. This fact implies that subscribers to these ISPs, particularly locating in Yangpyeong would have problems in accessing the Internet when malfunctions occur in peering arrangements. The detail shown in Figure 6.11 clearly shows the different levels of telecommunication possible between ISPs. The reason for this difference results from the number of interconnections with other ISPs in Internet Hubs. Enterprise IDC would be more susceptible to a weakened performance than the other ISPs since it peers with only three other ISPs (see Fig. 6.9).

[7] If the simulation follows the linkages failures to build reliability envelope, the number of computations dramatically increases since each r stage follows the complexity of $O(2^r)$ [r = the number of interconnections to be closed off] according to combinatoric rules.

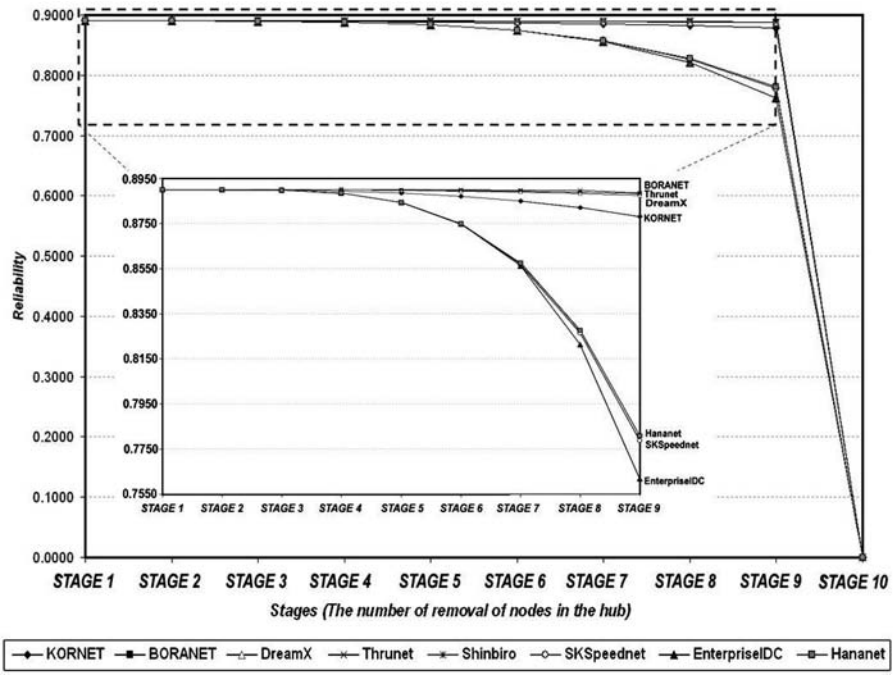


Fig. 6.11. Reliability envelope for the city pair Seoul-Yangpyeong

6.4.4 Reliability Variation to a Foreign Country

Finally, attention is paid to survivability of Korean Internets to be connected to the U.S., during potential hub nodes disruptions. As previously shown in Figure 6.9, seven ISPs have channels which connect to the United States. The Internet flow occurring in any domestic city nodes heading for the U.S. should pass through at least one of these foreign connection nodes. The main focus of this study was to examine the variations in reliability in terms of city nodes.

Figure 6.12 indicates how each city's average reliability is degraded as a result of the disruption of hub nodes. As an example of the worst scenario, shown in Figure 6.12(a), the inability to access the United States occurs only at the 7th stage, where all channels are unfortunately closed off. However, this extreme scenario is rare since only one combinatoric case out of 120 could actually cause this type of critical telecommunication failure. Most city nodes might suffer from unreliable access with lower reliability below 0.9, even at 5 stages. As a more probable scenario, Figure 6.12(b)

indicates that reliabilities of most city nodes are expected to experience a slight and gradual degradation by the 7th stage. That is, none of city nodes would maintain reliability above 0.9 after the 8th stage. Reliabilities decline significantly for all city nodes thereafter, as the number of zero-reliability cases would increase rapidly, highlighting that the total malfunction of telecommunication is realized as a plausible scenario. However, the actual impact in reality could be more severe than both scenarios if the capacity of linkages is not sufficient to treat the influx of the re-routed traffic derived from the disrupted hubs. Congestion in the network could hinder the propagation of routing information and appropriate peering refresh requests that cause the cascading failures (Curran et al. 2003).

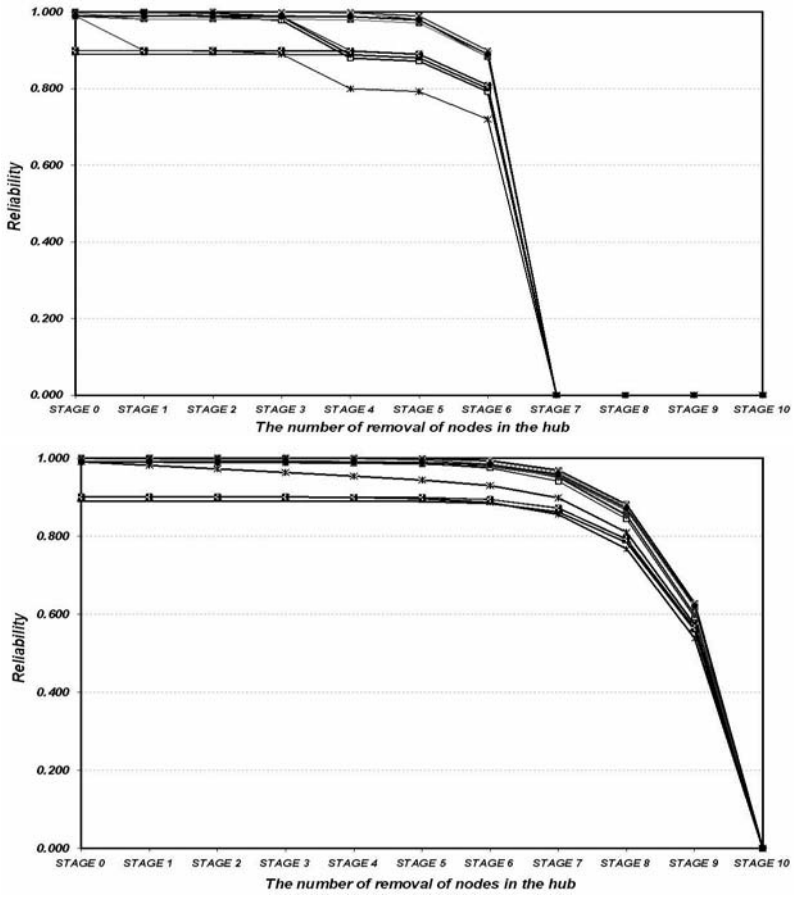


Fig. 6.12. Variations of lower bound reliability (a) and average reliability (b) of city nodes in Korea to the United States

6.5 Conclusions

This analysis examined the vulnerability of the Korean Internet in terms of reliability, taking into account the peering arrangements among 8 major ISPs. The contributions and findings of this study are summarized below.

In terms of methodology, this study introduces a probabilistic approach to assess the Internet as a critical infrastructure in the information era. Earlier work has shown that peering arrangements between Internet hubs can improve overall network performance by increasing network resilience. The concept of reliability developed here allows a more precise assessment of network vulnerability in contrast to previous measures. To illustrate the variation of reliabilities with increasing malfunctions, reliability envelopes are proposed in our work. Reliability envelopes can be utilized in examining the resiliency of hubs, city pairs and performance of ISPs by illustrating both the best and worst case scenarios possible. This study further indicates how to improve network reliability in terms of peering relationships.

A major finding of the paper is the significant role of Seoul in maintaining reliability for the rest of city nodes. The primacy of Seoul within the hierarchical urban system is also reflected both in the reliability analysis and in nodal accessibility. Seoul is regarded as one of the most vital locations in providing efficient Internet traffic exchange and better overall network performance by concentrating all peering relationships; however, at the same time, network failures in Seoul could cause a critical disaster, such as cascading failures of local networks that depend on major ISPs. Considering that Seoul is the main gateway for the Internet traffic between Korea and other countries, another hub should be implemented in order to ensure a desirable level of network reliability. As demonstrated in previous simulations of the U.S. Internet, the importance of distributing multiple Internet hubs geographically is essential for a more reliable network.

Several potential extensions from this study are suggested as future work. First, the resilience of the Internet can be reassessed again by considering more realistic situations with peering arrangements in multiple hubs. In addition, more realistic mathematical models can be suggested by embedding additional factors influencing survivability of the network into the model, such as link capacity of networks. Current studies only present a snapshot of a single time period. A longitudinal analysis from the initial stage to the current stage and beyond would help provide a better understanding of a country's development of the Internet.

References

Agrawal A and Barlow R E (1984) A survey of network reliability and domination theory *Operations Research* 32:478-492

Baran P (1964) On distributed communication networks *IEEE Transactions on Communications* 12(1):1–9

Ball M O, Golden B L and Vohra R V (1989) Finding the most vital arcs in a network *Operations Research Letters* 8:73–76

Ball M O, Colbourn C J, Provan J S (1995) *Handbooks in Operations Research and Management Science* Vol 7 Amsterdam and New York: Elsvier Science

Baybars I, Edahl R (1988) A heuristic method for facility planning in telecommunications networks with multiple alternate routes *Naval Research Logistics* 35:503-528

Broadband Week (2001) Tunnel vision?–train wreck sounds warning bell for fiber nets http://www.broadbandweek.com/news/010806/print/0100806_news_fiber.htm

Carreras B A, V E Lynch, I Dobson, and D E Newman (2002) Critical points and transitions in an electric power transmission model for cascading failure blackouts *Chaos* 12(4):985-994

Colbourn C J (1987) *The combinatorics of network reliability* New York: Oxford

Colbourn C J (1999) Reliability issues in telecommunications network planning In: Sanso, B., Soriano, P. (eds) *Telecommunications Network Planning* Boston: Kluwer Academic Press

Curran K, Woods D, Mcdermot N, and Bradly C (2003) The effects of badly behaved routers on Internet congestion *International Journal of Network Management* 13:83-94

Dodd Z. A (2002) *The Essential guide to telecommunications* 3rd (edn) New Jersey: Prentice Hall

Dotson W P, Gobien J O (1979) A new analysis technique for probabilistic graphs *IEEE Transactions on circuits and systems* 26(10):855-865

Doyle J C, D. Alderson D L, Li L, Low S, Roughan M, Shalunov S, Tanaka R and Willinger W (2005) The "robust yet fragile" nature of the Internet *Proceedings of the National Academy of Science of the United States of America* 102(41):14497-14502

Ellison R J, Fisher D, Linger R C, Lipson H F, Longstaff T, and Mead N R (1999) An approach to survivable systems http://www.cert.org/easel/nato1.doc

Fratta L and Montanari U G (1973) A Boolean algebra method for computing the terminal reliability in a communication network *IEEE, Trans Circuit Theory* 20:203-211

Gavish B and Neuman I (1992) Routing in a network with unreliable components *IEEE Transactions on Communications* 40(7):1248-1992

Gorman S P, Schintler L, Kulkarni R and Stough R (2004) The revenge of distance: Vulnerability analysis of critical information infrastructure *Journal of Contingencies and Crisis Management* 12(2):48-63

Grubesic and Murray A T (2006) Vital nodes, interconnected infrastructures, and the geographies of network survivability *Annals of the Association of American Geographers* 96(1):64-83

Grubesic T H, O'Kelly M E, Murray A T (2003) A geographic perspective on commercial Internet survivability *Telematics and Informatics* 20:51-69

Houck D J., Kim E, O'Reilly G P, Picklesimer D D and H. Uzunalioglu (2003) A network survivability model for critical national infrastructures *Bell Labs Technical Journal* 8(4):153-172

Huh W K, Kim H (2003) Information flows on the Internet of Korea *Journal of Urban Technology* 10:61-87

Huston G (1999) *The ISP survival guide* New York: John Wiley

International Telecommunication Union (2003) *ITU world telecommunication development report, Access Indicators for the information society* International Telecommunication Union

Internet Statistics Information System (2003) Internet connectivity map http://isis.nic.or.kr/english/sub03/sub03_index.html

Little R G (2002) Controlling cascading failure: Understanding the vulnerabilities of interconnected infrastructure *Journal of Urban Technology* 9(1):109-123

Malecki E J (2002) The economic geography of the Internet's infrastructure *Economic Geography* 78(4):399-424

Malecki E J, Gorman S P (2001) Maybe the death of distance, but not the end of geography: the Internet as a network in *Worlds of E-Commerce* (eds) T R Leinbach, S D Brunn New York: John Wiley 87-105

Moore M S, Pritsky N T, Riggs C and Southwick P V (2002) *Telecommunications: A beginner's guide* McGraw-Hill/Osborne

Moss M L, Townsend A (2000) The Internet backbone and the American metropolis *The Information Society Journal* 16(1):35-47

National Security Telecommunications Advisory Committee (NSTAC) (2003) *Internet peering security/Vulnerabilities task force report* The president's national security telecommunications advisory committee Washington D.C

National Computerization Agency of Korea (NCA) (2003 and 2004a) *Korea Internet white paper* Seoul: Korea: Ministry of Information and Communication

National Computerization Agency of Korea (2004b) *Broadband IT Korea Vision 2007: The Third master plan for Informatization Promotion* Seoul: Korea: Ministry of Information and Communication

National Internet Development Agency of Korea (2004) *Monthly Internet Statistics Report* KRNIC http://isis.nic.or.kr/report_DD_View/upload/rep200404.pdf

Norton W (2003) *Evolution of the U.S. peering ecosystem v1.1* Equinix white paper series http://www.equinix.com/pdf/whitepapers/PeeringEcosystem.pdf

O'Kelly M E, Kim H and Kim C J (2006) Internet reliability with realistic peering *Environment and Planning B: Planning and Design* 33:325-343

O'Kelly M E, Grubesic T H (2002) Backbone topology, access, and the commercial Internet 1997-2000 *Environment and Planning B: Planning and Design* 29:533-552

Shake T H, Hazzard B, Marquis D (1999) Assessing network infrastructure vulnerabilities to physical layer attacks" paper prepared for the National Information Systems Security Conference Arlington, VA, USA

Shier D R (1991) *Network reliability and algebraic structures* New York: Oxford

Shin S J, Correa H, Weiss M H (2002) A game theoretic modeling and analysis for Internet access market paper submitted at ITS-14 conference, pp.18-21 August, Seoul, Korea

Soni S, Gupta R and Pirkul H (1999) Survivable network design:The state of the art *Information Systems Frontiers* 1(3):303-315

Taaffe E J, Gauthier H L, O'Kelly M E (1996) *Geography of Transportation* Prentice Hall

Urban D and Keitt T (2001) Landscape connectivity: A graph-theoretic perspective Ecology 82(5):1205-1218

Wheeler D, O'Kelly M E (1999) Network topology and city accessibility of the commercial Internet *Professional Geographer* 51:327-339

White House (2003) The national strategy for the physical protection of critical infrastructures and key assets http://www.whitehouse.gov/pcipb/physical_strategy.pdf

Yoo Y B, Deo N (1988) A comparison of algorithms for terminal-pair reliability *IEEE Transactions on Reliability* 37(2):210-215

7 Railway Capacity and Train Delay Relationships

Lars-Göran Mattsson

Department of Transport and Economics, Royal Institute of Technology, SE-100 44 Stockholm, Sweden; Email: lgm@infra.kth.se

7.1 Introduction

Reliable transport infrastructure systems are vital for the functioning of modern societies. People in their everyday lives, as well as trade and industry, plan their activities on the assumption that it is possible to travel and to transport goods between different places in a fast, safe and predictable way. Over time the development of the transport infrastructure has allowed people and goods to be transported at higher speeds. This has contributed, for good or bad, to a spatial reorganisation of many human activities on a local as well as a global geographical scale. Transport systems, as all technical systems, are more or less reliable, however. This is an important aspect of the quality of transport services, which may have spatial implications. In many big cities capacity shortages lead to congestion and unreliable transports that hamper the development. In rural areas lack of alternative transport routes, in case the main route has to be closed for some reason, contributes to make these areas less attractive for location.

The railway system seems to be particularly vulnerable to various kinds of disruptive events. These events may have their origins within or outside the railway system related, for example, to technical faults, adverse weather and natural disasters. The tragic terror attacks on public transport passengers in Madrid 2004 and London 2005, show that not even intentional acts to hurt railway users can be excluded.

One reason to the railway system's lack of reliability is that it consists of many interacting subsystems, including the railway network with its signalling system, the electric power system, the rolling stock of engines and carriages, and the staff. All these subsystems work essentially in series. This makes the railway system particularly vulnerable to incidents. If one of the subsystem fails, or operates below its intended level of performance, the system as a whole will not function, or its level of service will be drastically reduced. Railways are also inherently inflexible because of the obvious restriction that trains must follow the tracks with no possibilities to switch track, unless there is a physical arrangement for it (Armstrong and McDonald 2005). Rail networks are also relatively sparse compared with road networks. This means that there are fewer possibilities to reroute rail traffic in case of disruptions.

Transport network reliability and vulnerability have been subject to considerable research in recent years (Bell and Ida 1997; Berdica 2002; Chen et al. 2002; Nicholson 2003; Morlok and Chang 2004). Various reliability concepts have been proposed and used such as connectivity reliability and terminal reliability, concerned with the probability of an existing path between two given nodes in a network, travel time reliability, which is the probability that the travel time between two given nodes is within a given time, and capacity reliability together with the related concept of capacity flexibility that are concerned with the ability of a transport network to accommodate the demand under varying conditions. Reliability studies are not only of interest in their own right. They can also contribute to vulnerability analysis. Components or parts of a system that are problematic from a reliability perspective may also be particularly critical from a vulnerability perspective.

The literature specifically devoted to quantitative reliability and vulnerability analysis of the railway system is rather sparse. It is mainly focused on travel time reliability. In particular, it deals with ways of modelling train delay and its relationship with capacity utilisation. This chapter will first introduce and discuss the concept of railway capacity and its connection with quality and reliability of train services. Then alternative methods of delay analysis will be reviewed including analytic, simulation and statistical approaches. In a final section some conclusions are drawn.

7.2 Quality of Train Services and Railway Capacity

Consider a train that runs between two stations. Let t_{min} be the minimum, or free, running time under ideal conditions, while t is the actual running

time. The difference between these two quantities $d = t - t_{min}$ could be called (total) delay. Part of the delay is already scheduled in the timetable as buffer or slack time. There are several reasons for this. Because of conflicts between different trains, it is seldom possible to schedule the trains so that they can all run according to their minimum running times. In addition, to reduce the risk for unscheduled delay d_u, some slack time is usually added to the running times, or to the durations of the stops in the timetable. The resulting difference between scheduled running time t_s according to the timetable and minimum running time t_{min}, is the slack time, which perhaps could be better termed scheduled delay, $d_s = t_s - t_{min}$. This means that (total) delay is the sum of scheduled and unscheduled delays $d = d_s + d_u(d_s)$, where the unscheduled delay varies from train to train in a stochastic way that may depend on the amount of scheduled delay for the own train d_s and for other trains. It may be observed that unscheduled delay may in fact be negative, if a train arrives at a station ahead of the timetable.

To summarise, the actual running time can be expressed as $t = t_{min} + d_s + d_u(d_s)$, that is the sum of minimum running time, scheduled delay and the resulting randomly occurring unscheduled delay. The sum of the first two of these components is the running time according to the timetable. If passengers would value travel time in the same way irrespectively of its origin, it would be natural to minimise the expected actual running time $E(t)$, when constructing a timetable. There is much evidence that passengers and carriers in fact put a much higher negative value on unscheduled delay than on the scheduled running time that is known to them in advance (for reviews see Bates et al. 2001; Wardman 2001). If it would be known how the unscheduled delay depend on the scheduled delay, as indicated by the function $d_u(d_s)$, the natural timetabling objective would rather be to find the scheduled delay d_s that minimises $t_{min} + d_s + \tau E(d_u(d_s))$. Here τ is the relative increase in the value of time for unscheduled delay compared with scheduled running time.[1] With increased scheduled delays for the trains utilising a line section, the expected unscheduled delays will typically go down. Whether it is worthwhile to increase the scheduled delays will depend on the value of τ and on the exact functional relationship between the scheduled delays and the expected unscheduled delays.

So far, the perspective has been on minimising overall delay to train movements. From the passengers' and carriers' perspective, minimising overall delay to passengers and goods is more relevant. Then different trains have to be weighted according to the number of passengers and to the value of delivering the goods on time.

[1] Arguably, to consider expected unscheduled delays only and not delay variations is also an oversimplification (see e.g. Bates et al. 2001).

Intuitively, a higher capacity utilisation leads to more delays of unexpected duration. How could capacity utilisation be measured? Capacity is an ambiguous concept with no commonly accepted definition. The general idea is to measure the maximum amount of traffic that a certain railway system, or a certain critical rail section, can accommodate in a given period of time. Burdett and Kozan (2006) suggest a precise definition of capacity "as the maximum number of trains that can traverse the entire railway or certain critical (bottleneck) sections(s) in a given duration of time". Capacity utilisation then "is the percentage time that an actual mix of trains utilises the section". They show how this definition can be translated into strict mathematical calculations under various circumstances.

The Swedish National Rail Administration has its own manual for calculating capacity utilisation, which is used in cost-benefit analyses of investments in the rail sector (see Banverket 2001). The basic idea is to determine the fraction of a specified time period that a certain line section between two stations is occupied by train paths.

Consider a single-track section between two stations without any possibilities for the trains to overtake or meet each other in between the stations. If the trains run alternately in the two directions, each train will occupy the whole track during its total running time between the stations. The total time the track is occupied is then the sum of the running times of all trains during the considered time period. In addition, some time has to be added to the total occupation time to allow the trains to meet at the stations. Let the total occupation time, or consumed capacity, be denoted T_s^o, where *s* indicates single track.

For a double-track section it is possible to separate the traffic by direction. The traffic will then be uni-directional on each track. Then a train does not occupy, or block, the track between the stations during its entire running time, since a consecutive train can be dispatched before it reaches the end station of the section. The occupation time is then the necessary time difference until the next train in the same direction can be dispatched. If the next train runs at the same or lower speed as the first one, this necessary time difference is the minimum headway time, partly determined by the subdivision of the track into block sections. If the second train runs at a higher speed, however, it is necessary to increase the time separation so that the second and faster train will not catch up with the first one. The necessary increase in the time separation is the difference in running times between the first (and slow) train and the second (and fast) one. This means that the total occupation time, or consumed capacity, for a double track with uni-directional traffic will be

$$T_d^o = \sum_{k=1}^{K} t_k^h + (t_k^r - t_{k+1}^r)^+ + t_k^p, \quad (1)$$

where $k = 1,2,...,K$, denotes the trains according to their timetable order during the considered period, t_k^h is the minimum headway, t_k^r is the running time, t_k^p is the additional time for a fast train to pass a slow train at a station, and where $(t_k^r - t_{k+1}^r)^+ \equiv t_k^r - t_{k+1}^r$ if $t_k^r > t_{k+1}^r$ and 0 otherwise.

The capacity utilisation is then calculated as

$$C_x = \frac{T_x^o}{aT}, \quad (2)$$

where x is s (single track) or d (double track), T is the length of the considered time period (typically 2 hours for the peak period and 18 or 22 hours for the whole day) and a is a reduction factor when calculating capacity utilisation; $a = 0.8$ when considering the whole day and $a = 1$ when considering the peak period. If the capacity utilisation is high enough, it could be expected that the railway services will be unreliable. Banverket (2001) indicates lack of capacity when capacity utilisation is above 80%.

It is interesting to observe that there are strong non-linearities in the provision of track capacity. Increasing the number of tracks on a line section increases the capacity of the section more than proportionate. This could be illustrated by some hypothetical examples. Assume that there is a line section between two stations that are 50 km apart. The trains are assumed to run at the same speed of 120 km/h. This means that the running time will be 25 min. If two trains meet at a station, each train is assumed to need a buffer time of 5 min.

First consider the situation with a single track. If the trains run alternately in the two directions, the theoretical capacity in terms of the maximum number of trains will be 2 trains per hour, since one train in each direction will occupy the line for 25 + 5 min. If instead the speed is 200 km/h, each train will occupy the line for 15 + 5 min, implying a theoretical capacity of 3 trains per hour.

One way of increasing the capacity of a single track would be to invest in meeting/passing stations. Assume that there is one such additional station in the middle of the railway section. If the speed is 120 km/h, a train will now occupy the track for 12.5 + 5 = 17.5 min rather than for 25 + 5 = 30 min. The theoretical capacity goes up from 2 to 3.4 trains per hour. If the speed instead is 200 km/h, the theoretical capacity goes up from 3 to 4.8 trains per hour. This way of increasing the capacity is probably quite sensitive to disturbances and leads also to longer scheduled running times

(including necessary buffers at the stations), however. The scheduled time will go up from 30 to 35 min for the lower speed level and from 20 to 25 for the higher speed level.

If the line section is served by a double track, the trains can be separated according to direction. If all trains run at the same speed, it will be the minimum headway between the trains that limits the capacity. The minimum headway depends on how the line section is subdivided into block sections and also on the speed of the trains. For the purpose of illustration, we simply assume a constant minimum headway of 5 min for all trains. Then the theoretical capacity per track goes up to 12 trains per hour instead of 2 to 4.8 trains per hour, depending on whether the comparison is made with slow or fast trains, and if there is an additional station in the middle of the line or not. It can also be noted that measures that would reduce the minimum headway, such as shorter block sections, or moving block sections, would further increase the capacity of a double-track section.

If for a double track there are trains running at different speeds, the theoretical capacity will be lower. Assume that the trains run alternately at 120 km/h and 200 km/h and that there is no additional station in the middle of the section. The running times will then be 25 and 15 min, respectively. Then it is necessary to add 10 min to the time the track is occupied for every second train, and in addition to that also 5 min for overtaking, see Eq. (1). The track will then be occupied during 25 min for every two trains (two minimum headways plus running time difference plus time for overtaking). This means that the theoretical capacity per track goes down to 4.8 trains per hour compared with 12 trains per hour for the situation with speed-homogenous trains. With an additional station in the middle of the section the theoretical capacity increases to 6 trains per hour and track, because the running time difference between the trains will then go down to 5 min.

If the number of tracks is expanded to four, the trains could be separated both with respect to direction and speed. Then the theoretical capacity is again determined by the minimum headway. The theoretical capacity per track will be 12 trains per hour. This holds both with and without an additional station and irrespectively of the speed.

These simple examples of capacity calculation are summarised in Table 7.1. They illustrate several interesting properties of the provision of capacity. If, for the sake of illustration, the average capacity per track is calculated for one, two and four tracks, these values are 3.3, 8.7 and 12 trains per hour, respectively. This indicates that the capacity per track is strongly non-linear in the number of tracks. Assuming constant or decreasing marginal track investment cost, railway capacity exhibits large economies of scale. The table also shows that the capacity does not only depend on

properties related to the technical infrastructure in form of the number of tracks and meeting/passing stations. The capacity also depends on the average speed of the trains and on the speed heterogeneity. Speed heterogeneity, not the least, is very detrimental to the achieved theoretical capacity. By reducing the speed of the fast trains, it is in fact possible to increase the capacity considerably, for example from 4.8 to 12 trains per track and hour as is seen by comparing rows 7 and 5 in the table.

Table 7.1. Theoretical capacity per track by number of tracks, speed, speed heterogeneity and presence of additional station

Number of tracks	Speed [km/h]	Additional station	Capacity per track [trains/h]
1	120	No	2
1	200	No	3
1	120	Yes	3.4
1	200	Yes	4.8
2	120	No/yes	12
2	200	No/yes	12
2	120/200	No	4.8
2	120/200	Yes	6
4	120/200	No/yes	12

Finally, it could be noted that rail capacity follows the fundamental law of traffic that is well-known from road traffic – flow is equal to speed times density (number of trains per unit of track length) $f = v \cdot d$ (see, e.g., Bella and Ida 1997, p. 68). Since theoretical capacity c is the maximum flow, this leads to $c = f_{\max} = \max(v \cdot d) = \hat{v} \cdot \hat{d}$, where $\hat{v}$ and $\hat{d}$ are the values of speed and density that together simultaneously maximise flow.

The application of this formula is illustrated by two examples. Consider first the situation with one track according to the first row in Table 7.1. Then there is at most one train at the time on a 50 km long railway section, that is $\hat{d} = 1/50$ trains/km/track. Since the average speed is $\hat{v} = 100$ km/h, when adjusting for the time to meet at the station, the capacity is $c = 100/50 = 2$ trains/h/track. When the trains run in one direction on each track and at the same speed, the capacity is calculated under the assumption of a constant headway. For the assumed headway of 5 min, the average density is $\hat{d} = 60/(5\hat{v})$ and hence the capacity is $c = \hat{v} \cdot 60/(5\hat{v}) = 12$ trains/hour/track. This calculation is valid for trains running at the same speed independent of the actual level of speed and hence for rows 5, 6 and 9 in Table 7.1. Increasing the number of tracks from one to two is radically to increase the maximum possible density. This is why there are such large economies of scale. The introduction of more stations also allows for a

higher maximum density, in that case partially at the expense of a lower average maximum speed due to more frequent stops for meeting or overtaking. The reason why speed heterogeneity has such a negative effect on capacity is that the average maximum density is radically reduced, given that the timetable is designed so that faster trains never will catch up with slower trains. By letting fast and slow trains run on separate tracks, which is possible when there are four tracks, the trains can run at their maximum speed without having to lower the density. This is why there is also an economies-of-scale effect, when the number of tracks is increased from two to four.

To summarise, capacity can be defined as the maximum number of trains that can traverse a railway section per time unit. The capacity depends on the technical infrastructure in terms of the number of tracks, the number and location of stations where the trains can meet or overtake each other, and the subdivision of the line section into block sections. The capacity also depends on conditions related to the character of the train service such as the average speed, the speed heterogeneity and in which order different trains are scheduled. Capacity is a kind of theoretical indicator that measures the maximum number of trains under ideal conditions, when all trains can run exactly according to the timetable. In reality there will be many kinds of disturbances. These disturbances will lead to unscheduled delays that will make the service more or less reliable. It could be hypothesised that the degree of unreliability, or the risk for delays, will depend on the capacity utilisation. We will continue by reviewing how such relationships can be derived.

7.3 Analytic Methods of Delay Analysis

There are some quite interesting ways of analysing train service reliability by analytic methods. Compared with simulation methods, analytic methods are usually much faster to apply, which make them particularly useful at a strategic planning and design stage, when the future timetable is still unknown, and hence many variants are conceivable. To be able to derive an analytic model, it is usually necessary to make a number of simplifying assumptions. This drawback is less of problem at a planning stage when many design parameters, operating conditions and the future demand anyhow are very uncertain.

To be able to relate reliability of train services to capacity utilisation, useful indicators of reliability are needed. Carey (1999) presents an insightful analysis of the mechanism behind delays. He separates between

exogenous, or primary, delays and knock-on, or secondary delays. An exogenous delay is caused by some external event. It could be a breakdown or a failure of equipment or infrastructure, an extended stop at a station because boarding and alighting take more time than expected, and it could also be a train or a crew that is late to the starting position. Knock-on delays are those delays that occur in interaction between an exogenously delayed train and the other trains that are scheduled in the timetable. If a train arrives late to a station because of some exogenous delay, this may in turn delay other trains. To what extent this happens depends on the timetable and on the amount of buffer time in it. Under high capacity utilisation, one delayed train can cause delays to several other trains over a large area and a long period of time. By reducing capacity utilisation it is possible to decrease the risk for knock-on delays and hence to improve the reliability of the timetable and also to restrict the area and the time period that might be affected. By the design of the timetable it is thus possible to reduce the consequences of exogenous delays. To reduce the exogenous delays on the other hand, it is necessary to reduce the probabilities for the events that are causing these delays. The exogenous delays are normally not possible to affect by the construction of the timetable.

Carey's (1999) analytic approach is relevant to double track services. Trains running in opposite directions are then running on different tracks. He derives formulas for calculating probability density functions of knock-on delays as a function of the amount of headways present in the timetable and given probability density functions for exogenous delays. As an alternative, he also derives formulas for the calculation of expected knock-on delays. He also discusses some heuristic measures that do not require probability information on exogenous delays.

De Kort et al. (2003) present an interesting approach to determine the capacity of a planned railway infrastructure layout under uncertainties. They develop a methodology to calculate the maximum number of train movements that can be executed on a particular infra-element in a given time period with probability greater than or equal to p. The threshold p can be considered a requirement on the reliability of the service. Their approach allows the analyst to make explicit assumptions on travel time uncertainties based on, for instance, historically observed uncertainties. They apply their approach to a planned high-speed double-track line in the Netherlands. The line is slightly more than 100 km long and involves three tunnels, of which the longest is about 7 km. Each tunnel has a separate tunnel tube per direction. For evacuation reasons only one train at a time is allowed in each tunnel, so that the non-occupied tube can be used for escape in case of emergency. The tunnels then function as single-track bottlenecks on an essentially double-track line. From the start 8 trains per

hour and direction could be expected. This volume is supposed to go up to 16 trains per hour by 2015. The question is if the capacity of the line is large enough considering the bottlenecks created by the tunnels. Table 7.2 shows the capacity of the longest tunnel according to the analysis by de Kort et al. (2003). Apparently the capacity of the proposed layout will be far below assumed demand for any reasonable level of the reliability requirement p. The approach suggested by de Kort et al. is a way of indirectly determining a causal relationship between traffic volume and reliability of the service. It can be noted that the capacity decreases rapidly for a high requirement on the reliability. This study also illustrates how short bottlenecks will reduce the overall reliability for a long line section.

Table 7.2. Capacity of the longest tunnel by the reliability requirement p. Source: de Kort et al. (2003)

Reliability requirement p	0.70	0.75	0.80	0.85	0.90	0.95	0.99
Capacity [trains per hour and direction]	5	5	5	4	4	4	2

Huisman and Boucherie (2001) develop a very elegant model for the analysis of delay on a double-track section with heterogeneous traffic, and where overtaking is not possible or allowed. The aim of the model is to forecast secondary, or knock-on, delay as a function of primary, exogenous, delay. The strength of the model is that it can handle the investment case when there are only forecasts about the frequency of different types of trains but no detailed timetable. It can also handle cyclic timetables, i.e., when the timetable is the same for every hour. In the latter case it can be used to study in which order the trains should be scheduled to minimise overall delay.

The mathematical idea of the model is that the running time R_n of train n in a sequence of trains satisfies the recurrence relation

$$R_0 = F_0, \tag{3}$$

$$R_n = \max(F_n, R_{n-1} - A_n), \tag{4}$$

where F_n is the free running time of train n that could be actual if this train would not be delayed by other trains and $A_n = T_n - (T_{n-1} + H_n)$ is the actual buffer time between the trains, where T_n is the time train n enters the railway section and H_n is the minimum headway. There is a simple intuition behind this recurrence relationship. The running time R_n of train n can never be shorter than its free running time F_n. In addition, it must be so long that train n does not arrive earlier to the end point than the arrival time of the previous train plus the minimum headway, $T_n + R_n \geq T_{n-1} + R_{n-1} + H_n$. By making suitable statistical assumptions for

the distributions of the free running time F_n, and the actual buffer time A_n, it is possible to derive the steady-state distribution for the running time R_n. How the frequency and speed distribution of the trains affect the delays, that is the differences between actual running times and free running times, can then be studied.

The use of the model will only be illustrated for long-term analysis. Then, typically, there is no timetable available. The only information is about the frequency of different train types. Huisman and Boucherie (2001) consider in a case study a 67 km long double-track section in the Netherlands. There are three train types, regional, interregional and intercity trains, each characterised by a different free running time. The free running times are assumed to be deterministic amounting to 33, 36 and 48 min for intercity, interregional and regional trains, respectively. One simplifying assumption that is reasonable when the timetable is not known is that trains of different types arrive in a random order. All train types are assumed to be equally frequent on the average. The buffer time is assumed to be exponentially distributed and the minimum headway is set to 2 min.

Fig. 7.1 shows how the mean delays, and the delay probabilities, vary with the number of trains per hour and track. Regional trains are not shown, because under the assumptions made, they always run at the deterministic free running time of 48 min. All delays in this example are secondary because of the assumption of deterministic free running times. When the number of trains is very low, there are practically no delays. A faster train is almost never caught behind a slower one. When the number of trains increases, the mean delays go up as well as the delay probabilities. When the total number of trains reaches 30 per hour and track, which is the maximum that could be achieved given the minimum headway of 2 min, all trains run at the speed of the regional trains and the mean delays are 48 – 33 = 15 and 48 – 36 = 12 min for intercity and interregional trains, respectively.

It should be noted that if 30 trains per hour is considered to be the capacity of the track, this is in fact another definition of capacity than the one employed by Banverket (2001). In the latter case, capacity is defined as the maximum number of trains per time unit given that the trains are allowed to run according to their scheduled running times in the timetable. That level of capacity can be exceeded at the price of increasing levels of delay for fast-running trains. This is what Fig. 7.1 illustrates. The definition of capacity applied in Huisman and Boucherie (2001) is the maximum number of trains that could run on the track per time unit given the minimum headway of 2 min. This is more of a physical capacity measure determined by the technical infrastructure, including the signalling system and the subdivision of the track into block sections.

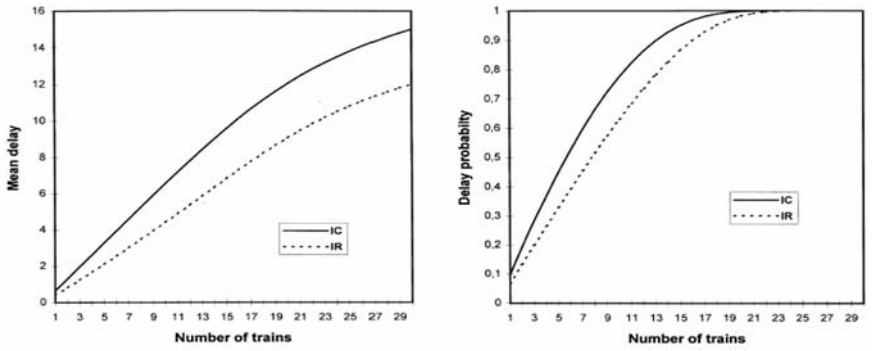


Fig. 7.1. Mean delays and delay probabilities as a function of the number of trains per hour and track. Source: Huisman and Boucherie (2001)

A limitation with the previous model is that it is only applicable to railway sections with no stations for meeting and overtaking. In a companion paper Huisman et al. (2002) propose a network queuing model by which the performance of a total railway network system of double tracks can be evaluated.

There is also a large literature on optimal scheduling of train operations (e.g. Higgins et al. 1996; Şahin 1999; Dorfman and Medanic 2004; Ghoseiri et al. 2004). Such optimisation models can be applied in decision support systems to help train dispatchers to reschedule trains in real time when delays already have occurred. They can also be used for the design of an optimal timetable or for the analysis of impacts of timetable changes or railway infrastructure changes. These optimisation problems can be quite time-consuming to solve. Usually it is necessary to rely on heuristic techniques for integer and non-linear optimisation.

7.4 Micro-Simulation Methods of Delay Analysis

How trains interact with each other on a real railway line and how a primary, or exogenous, delay originally affecting one train may interfere with the operations of other trains and hence cause secondary, or reactionary, delays, is a very complex process. This process depends on a number of parameters related to the design of the track system, the signalling system and the rolling stock equipment. Simulation is the main avenue currently available to model this process in greater detail. A number of such models have been developed in various places. No attempt will be made to review the full spectrum of such models. Rather a few studies will be discussed to

illustrate the use of simulation models to study the relationship between capacity utilisation and delays.

In a study by Hallowell and Harker (1998) the main focus is on examining how a previously developed model for delay calculation can be used as a tool for optimising train schedules. To this end a number of Monte Carlo simulations of two railway lines in the US are performed. A base case is compared with a case, where the schedules are optimised by means of the developed model. The results for only one of the lines in this study will be discussed here. This particular line is 444 km long and has less than 3% double track. The traffic volume scenarios low, medium and high correspond to 22, 29 and 33 trains per day. The uncertainty in the departure time of the trains at the origin of the railway line is exogenously given. The low, average and high scenarios have 50, 100 and 150%, respectively, of the average standard deviation (SD) of departure time error of the line.

The results from the simulations are displayed in Table 7.3. Delay is here defined as the difference between the actual and the free (unrestricted) running time of a train. Mean delay is obviously quite sensitive to traffic volume, while fairly insensitive to departure time uncertainty. For instance, for the average level of departure uncertainty, mean delay increases from 51 to 90 min, as traffic volume increases by 50% from the low to the high scenario. For the standard deviation of delay, the situation is rather the other way round. The standard deviation is rather sensitive to departure uncertainty, while not so sensitive to traffic volume. For medium traffic volume, the standard deviation goes up from 18 to 23 min, as departure uncertainty increases from low to high. For average departure uncertainty, standard deviation goes only up from 18 to 20 min, as traffic volume increases from low to high.

Table 7.3. The dependence of mean delay and standard deviation of delay on traffic volume and departure uncertainty for a 444 km long railway line in the US. Source: Railway line 1 in Hallowell and Harker (1998)

Traffic volume	Low			Medium			High		
Departure uncertainty	Low	Avg	High	Low	Avg	High	Low	Avg	High
SD departure error (min)	2.6	5.1	7.7	2.6	5.1	7.7	2.6	5.2	7.7
Mean delay (min)	49.7	51.2	51.1	72.1	71.0	74.1	89.5	89.7	89.7
SD delay (min)	12.6	17.9	22.3	18.0	19.7	23.2	16.3	19.7	24.0

The work by Rietveld et al. (2001) does not represent a typical simulation study. It is quite interesting, however, because it looks at the issue of travel time unreliability of public transport from a passenger (or demand) perspective rather than from a train service (or supply) perspective. This means a shift in attention from the reliability of train arrival times to the

reliability of passenger arrival times. One obvious consequence is that it is necessary to weight train delays with the number of affected passengers. But more importantly, there is also a shift in attention from the reliability of single trains to the reliability of complete trip chains of public transport users. This means that the risk of missed connections will come into focus.

The same authors also suggest different measures of reliability and a methodology for calculating reliability for journeys or trip chains that may include transfers between different vehicles or modes of transport. To apply this methodology, estimations of density functions of departure and travel times for different modes of transport are needed. Then by drawing a representative sample of public transport chains from some population, it is possible to compare for each chain the scheduled arrival time with the simulated arrival time based on the estimated density functions. In calculating the simulated arrival times, it is assumed that different time elements in the chain are statistically independent. Various measures of reliability can then be calculated. It is also possible to study spatial variations in the occurrence of delays.

Table 7.4 displays the average scheduled and simulated travel times for selected types of public transport chains sampled from the Dutch annual travel survey for morning peak, off-peak and Sundays, respectively. Somewhat surprisingly, the period of the day seems to have little impact on the difference between average scheduled and simulated travel time. The number of transfers is more decisive.

Table 7.4. Average scheduled (min) and average simulated (as % of scheduled) travel times for some selected types of public transport chains sampled from the Dutch annual travel survey. Source: Rietveld et al. (2001)

	Average scheduled travel time			Average simulated travel time as a % of the scheduled travel time		
	Morning peak	Off-peak	Sundays	Morning peak	Off-peak	Sundays
Bus	28.8	28.7	28.1	100.3	102.4	101.1
Bus/bus	50.5	54.5	55.2	106.1	106.4	110.9
Train/bus	53.7	60.7	65.3	109.5	103.6	109.6
Bus/train	58.0	55.2	64.6	108.8	109.2	102.3
Bus/train/bus	74.3	71.0	85.5	111.6	111.8	110.5
All chains	59.9	61.8	65.1	110.0	109.9	111.4

This methodology can be extended to allow different policies to improve reliability to be evaluated. In the particular Dutch study, it turned out that enhancing the use of bicycle as an entrance and exit mode was a promising policy of increasing the reliability of the public transport system.

In recent years in Sweden there has been a number of applications of a

commercial microsimulation system, RailSys, developed by the Institute for Transport, Railway Construction and Operation at the University of Hannover in Germany.[2] RailSys is originally developed for timetable construction and planning concerning new or existing lines, nodes and networks. By simulation of non-disrupted and disrupted operations it can also be used to judge the stability or quality of a timetable.

The use of RailSys will be illustrated by relating its application to a vulnerability study of major, or exceptional, disturbances in the operations of rail services (see Wiklund 2003). This application is an attempt to simulate the consequences of a serious breakdown of a technical system supporting the railway operations. The specific breakdown was a fire in the interlocking system at the railway station, Järna, situated on a major double-track railway line some 40 km southwest of Stockholm. The breakdown occurred in the summer of year 2000. As a consequence of this fire all signals at the station were put out, and it was not possible to use the switches. When the traffic temporarily could start again, the number of trains was reduced, no switches were used, the maximum speed through the station was set down to 40 km/h and a manual signalling system replaced the electric signals. The station then essentially functioned as a double track with reduced maximum speed. To apply RailSys to this situation it was necessary to carry out an extensive calibration of the model. In the end, the simulated average delay was 16.1 min, compared with the observed average delay for the first ten days after the breakdown that was 17.6 min.

Fig. 7.2 displays the simulated time profile of the average delay for three levels of traffic volume: the same number (66 trains per day), more trains (84 per day) and fewer trains (50 per day) than were actually run after the breakdown. The average delay varies quite substantially over the day with peaks around 10 am and 8 pm. For the level of traffic volume that was actually carried out, the average delay goes up to 30 min in the morning peak and to 45 min in the evening peak. The higher level of traffic volume corresponds to the actual volume before the breakdown, with the exception of commuter trains. Had this level been kept after the breakdown, the average delay would have gone up to 46 min in the morning peak and to 70 min in the evening peak. For the lower level of traffic, maximum average delay is 30 min with peaks both in the morning and in the early evening. It should be remembered that all values are average delays. The actual delays will vary considerably around these values, which is more clearly illustrated in the next figure.

Fig. 7.3 summarises the results in form of average delay and standard deviation over 24 hours as a function of the number of trains per hour. Al-

[2] See the website http://www.ive.uni-hannover.de/software/railsys/info_en.shtml

though the general tendency is that the average delay increases with the number of trains, the relationship is not monotonic. The explanation may be that when the number of trains increases, it is sometimes necessary to increase the scheduled running time to make the timetable feasible. Some buffer time is then introduced in the timetable that reduces the risk of delays. There are considerable random variations around the average delay. The actual running time will be rather unstable, and increasingly so as the capacity utilisation goes up.

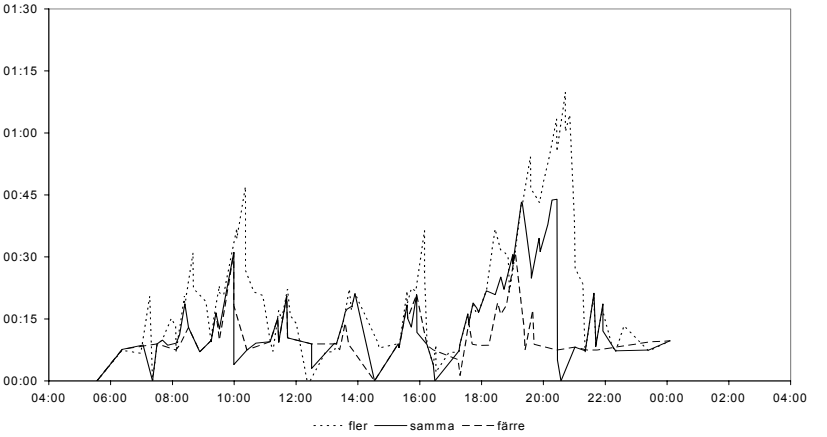


Fig. 7.2. Simulation of the burnt down of the interlocking system at Järna station. The horizontal axis represents the time of the day and the vertical axis the average delay (h:min). Three levels of capacity utilisation are presented: "fler" indicates more trains, i.e. 84 trains per day; "samma" indicates the same number of trains, i.e. 66 trains per day; "färre" indicates fewer trains than were actually run after the breakdown, i.e. 50 trains per hour. Source: Wiklund (2003)

7.5 Statistical Analysis of Delays

A third way of studying the relationship between capacity utilisation and delay is by statistical analysis. There are few such studies reported in the literature (for a review, see Olsson and Haugland 2004). One exception is the pioneering study by Gibson et al. (2002) based on data from the UK rail network after the introduction of access charges. They develop a method for defining the marginal cost of an additional train (the "congestion cost") including the extra costs for the train operators and the passengers. This congestion cost reflects the costs of reactionary delays caused by a given level of exogenous delays.

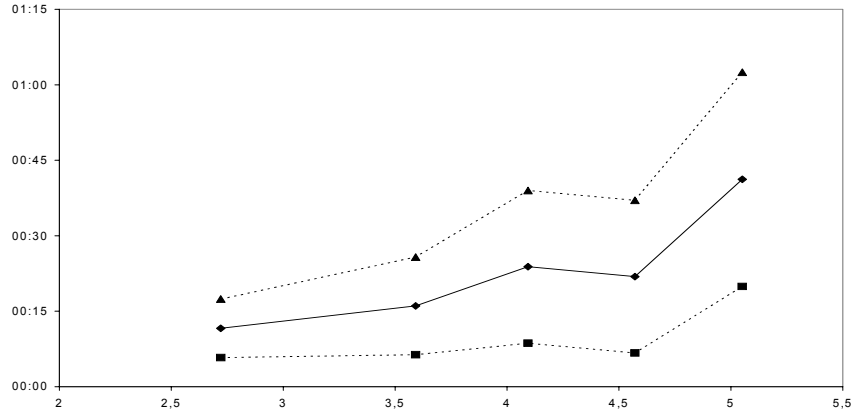


Fig. 7.3. Simulation of the burnt down of the interlocking system at Järna station. The horizontal axis represents trains per hour and the vertical axis the average delay (h:min) over 24 hours. The unbroken line represents average delay whereas the dotted lines indicate ± one standard deviation. Source: Wiklund (2003)

The purpose of the study by Gibson et al. (2002) is to relate reactionary delays to capacity utilisation to enable congestion costs to be properly reflected in access charges.[3] To this end they regress reactionary delays against capacity utilisation for different track sections in the rail network and for different time periods. They find that an exponential relationship best fits the data:

$$D_{it} = A_i \exp(\beta C_{it}), \tag{5}$$

where D_{it} is the reactionary delay on track section *i* in time period *t*, A_i is a section specific constant, β is a route specific constant, and C_{it} is the capacity utilisation on section *i* in time period *t*. They are able to find positive and statistically significant relationships for 20 out of 24 routes across the network. The β-value typically varies between 1.1 and 4.1.

This study shows that it is possible to derive significant relationships between reactionary delays and capacity utilisation. The estimated constants, A_i and β, reflect spatial differences including "normal" amount of exogenous delays for different track sections. This is both a strong and a weak point of the approach. The relationships are empirically based but do not reveal how exogenous delays cause reactionary delays. Hence they are not useful for predicting the effects of an altered level of exogenous delays.

[3] Kozan and Burdett (2005) suggest alternative methodologies for rail access charging.

7.6 Final Discussion and Conclusions

This chapter aims at reviewing some possibilities of analysing train delays and their relationships with capacity utilisation. As discussed in Section 7.2 (total) delay could be defined as the difference between actual and minimum running time under ideal conditions. The delay could then be divided up into scheduled delay, that is the difference between scheduled running time and minimum running time, and the rest that could be termed unscheduled delay. The reason to include scheduled delay in the definition is that scheduled delay is a kind of policy variable that will affect unscheduled delays through its effect on capacity utilisation. From a policy perspective it would be natural to choose the scheduled delay so as to optimise a weighted sum of the different components of the delay. There is ample evidence that travellers attach different monetary values to scheduled and unscheduled delays (see Bates et al. 2001; Wardman 2001). Since the relationship between scheduled delay and unscheduled delay is stochastic and highly non-linear, and since there are also operational costs that depend on the choice of scheduled delay, this optimisation is far from trivial. By including the scheduled delay in the definition of the delay, the trade-off between scheduled and unscheduled delay is at least made explicit.

Unscheduled delays consist of two parts of different origin, primary or exogenous delays and secondary or reactionary delays. This distinction is very important from an analytic point of view. A primary delay is caused by some exogenous event and is by definition independent of capacity utilisation. Secondary delays are a resulting effect of a primary delay. If a train has to stop or reduce its speed below what is assumed in the timetable for some reason related to a primary delay, trains that follow close after this train may also have to stop or drive slower. If the line has only one track, meeting trains may have to wait longer and at different places than scheduled in the timetable. This will incur secondary delays on these trains as well. The amount of secondary delays does not only depend on the frequency and duration of primary delays but also on the capacity utilisation. The higher the capacity utilisation is, the more likely it is that primary delays will knock-on secondary delays on following or meeting trains.

Although primary delays are independent of capacity utilisation by definition, it may not be so simple to verify this in an empirical study because of the problem of classifying delays correctly. Interestingly, Gibson et al. (2002) are able to do that in their study.

It is probably extremely difficult to derive a causal model that could explain the occurrence of the events that will lead to primary delays. How-

ever, by statistical regression it is possible to estimate equations by which the amount of primary delays can be predicted based on factors such as maintenance status of the track and the rolling stock, maintenance activities performed, traffic volumes, shortage of staff and weather conditions. Although such analyses are quite straightforward in principle, very few are reported in the literature (Olsson and Haugland 2004).

Most of the literature that has been covered in this chapter deals with the impacts of primary delays and capacity utilisation on secondary delays. Three different approaches are discernible: analytic methods, micro-simulation methods and statistical analyses based on empirical data. All three approaches have their advantages and disadvantages.

The analytic methods are often using elements from queuing theory and optimisation that may lead to mathematical problems that could be time-consuming to solve. They do not usually require so much input data as simulation models. In particular, they can in many cases be applied in a meaningful way without knowing the exact timetable. This makes them useful for strategic decisions – especially in a sketch-planning phase when many investment alternatives need to be evaluated, and when it is unclear how the new infrastructure actually will be used in the future. Analytic models usually apply some form of simplifying assumptions to make them mathematically tractable. This has the drawback that the quantitative conclusions can be less precise and reliable. It may also be difficult to validate the assumptions on which the model is based.

Simulation models offer the most detailed representation of a railway system. In fact, micro-simulation is today the only reasonable way to model in any detail the very complex processes by which different trains interact with each other and with the infrastructure. The other side of the coin is that they require very detailed data about the infrastructure, the performance of the trains and, perhaps most importantly, about the timetable.

Many of the simulation models that are available on the market are primarily tools for timetabling. When facing a strategic decision about future investments, the exact timetable is not known. It is then necessary to make some assumptions that will introduce uncertainties into the results irrespectively of how well the model represents the railway system. The increased preciseness with simulation models may therefore be somewhat illusory. Since the simulation models available on the market are commercial products, they are not always particularly transparent with respect to underlying assumptions. On the other hand, they do not require any specific mathematical skill for their implementation.

As has been illustrated in Section 7.4, it is possible to use a simulation tool like RailSys to study how primary delays under varying capacity utilisation will cause secondary delays. Such analyses seem to be the most pre-

cise way of deriving causal relationships for secondary delays. One problem is that each specific case has to be simulated with all its details. This is time-consuming and may also disguise general tendencies in the results from all details.

Statistical analyses of empirical data seem to be the only realistic way of modelling the occurrence of primary delays. The study by Gibson et al. (2002) shows that regression analysis could be a useful method for establishing empirical relationships between capacity utilisation and secondary delays, given the prevailing level of primary delays. Their analysis was explicitly designed to provide information about congestions costs to the manager of the railway infrastructure in the UK for determining access charges to the operators.

Most delay studies deal with delay to train movements, i.e. they apply a supply, or producer, perspective. From a welfare perspective a demand, or consumer, perspective would be more interesting. What delays will passengers or goods suffer from during their movement from the origins to their final destinations? This is far more complicated to study, since it is then also necessary to model the transfers between different modes or lines in the transport chains. Rietveld et al. (2001), however, offer an interesting approach to this issue.

Acknowledgments

The author is grateful for very helpful comments and suggestions from the editors and from two anonymous referees. The research was given financial support from the Swedish Institute for Transport and Communications Analysis. An earlier version of the paper was presented at the 1st International Seminar on Railway Operations Modelling and Analysis in Delft 2005.

References

Armstrong J and McDonald M (2005) The vulnerability of railway operations to disruptive incidents, and a proposed framework for addressing the issue. Paper presented at the 1st International Seminar on Railway Operations Modelling and Analysis (RailDelft2005), Delft.

Banverket (2001) Beräkningshandledning: Hjälpmedel för samhällsekonomiska bedömningar inom järnvägssektorn (Calculation manual: Aid for cost-benefit analysis in the railway sector). Handbok BVH 706.00, Banverket, Borlänge (In Swedish)

Bates J, Polak J, Jones P, Cook A (2001) The valuation of reliability for personal travel. Transportation Research E 37: 191–229

Bell MGH and Iida Y (1997) Transportation network analysis, Wiley, Chichester

Berdica, K (2002) An introduction to road vulnerability: what has been done, is done and should be done. Transport Policy, 9: 117–127

Burdett RL, Kozan E (2006) Techniques for absolute capacity determination in railways. Transportation Research B, in press

Carey M (1999) Ex ante heuristic measures of schedule reliability. Transportation Research B 33: 473–494

Chen A, Yang H, Lo HK and Tang WH (2002) Capacity reliability of a road network: an assessment methodology and numerical results. Transportation Research B 36: 225–252

De Kort AF, Heidergott B, Ayhan H (2003) A probabilistic (max, +) approach for determining railway infrastructure capacity. European Journal of Operational Research 148: 644–661

Dorfman MJ, Medanic J (2004) Scheduling trains on a railway network using a discrete event model of railway traffic. Transportation Research B 38: 81–98

Ghoseiri K, Szidarovszky F, Asgharpour MJ (2004) A multi-objective scheduling model and solution. Transportation Research B 38: 927–952

Gibson S, Cooper G, Ball B (2002) Developments in transport policy: The evolution of capacity charges on the UK rail network. Journal of Transport Economics and Policy 36: 341–354

Hallowell SF, Harker PT (1998) Predicting on-time performance in scheduled railroad operations: methodology and application to train scheduling. Transportation Research A 32: 279–295

Higgins A, Kozan E, Ferreira L (1996) Optimal scheduling of trains on a single line track. Transportation Research B 30: 147–161

Huisman T, Boucherie RJ (2001) Running times on railway sections with heterogeneous train traffic. Transportation Research B 35: 271-292

Huisman T, Boucherie RJ, van Dijk NM (2002) A solvable queueing network model for railway networks and its validation and application for the Netherlands. European Journal of Operational Research 142: 30-51

Kozan E, Burdett R (2005) A railway capacity determination model and rail access charging methodologies. Transportation Planning and Technology 28: 27– 45

Morlok EK, Chang DJ (2004) Measuring capacity flexibility of a transportation system, Transportation Research A 38: 405-420

Nicholson A (2003) Transport network reliability measurement and analysis. Transportes XI(2): 49–62

Olsson NOE, Haugland H (2004) Influencing factors on train punctuality: results from some Norwegian studies. Transport Policy 11: 387–397

Rietveld P, Bruinsma FR, van Vuuren DJ (2001) Coping with unreliability in public transport chains: A case study for Netherlands. Transportation Research A 35: 539-559

Şahin İ (1999) Railway traffic control and train scheduling based on inter-train conflict management. Transportation Research B 33: 511-534

Wardman M (2001) A review of British evidence on time and service quality valuations. Transportation Research E 37: 107-128

Wiklund M (2003) Allvarliga funktionsstörningar i baninfrastrukturen: Beräkning av effekter på tågtrafiken (Serious breakdowns in the track infrastructure: calculation of erffects on rail traffic). VTI meddelande 959 (In Swedish)

8 A Reliability-based User Equilibrium Model for Traffic Assignment

William H.K. Lam[1], Ning Zhang[2] and Hong K. Lo[3]

[1]Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hong Kong
[2]School of Economics and Management, BeiHang University, Beijing
[3]Department of Civil Engineering, Hong Kong University of Science and Technology, Hong Kong

8.1 Introduction

One of the major problems encountered by traffic professionals is how to predict driver route choices in congested and unreliable road networks. Indeed, traffic flow prediction is of fundamental importance for strategic transport planning, network design and long-term road improvement problems. A new reliability-based user equilibrium (RUE) model is proposed in this Chapter, premising on the fact that drivers consider both travel time and reliability in their route choices (Lam and Small, 2001).

Conventional approaches for traffic assignment consist of the user equilibrium (UE) principle, as well as the system optimal (SO) and stochastic user equilibrium (SUE) principles. The UE principle, also referred to as Wardrop's first principle, states that the travel times on all used routes per origin-destination (OD) pair are equal and minimal; whereas those on unused routes are equal or higher (Wardrop, 1952; Dafermos and Sparrow, 1968; Smith, 1979). The SO principle, also referred to as Wardrop's second principle, defines the traffic assignment pattern such that the total network travel time is minimized. The SUE principle (Daganzo and Sheffi, 1977; Sheffi and Powell, 1982) characterizes the traffic flow pattern such that the probability of the minimal perceived travel time/cost on a path equals the path choice proportion over the travel demand. All of these models, however, do not explicitly consider the fluctuation of link travel times due to minute-to-minute variations in link flows within the same hourly period. In

a multi-lane freeway, these traffic flow variations are mainly due to the fact that drivers tend to increase their speed variations as the speeds between adjacent lanes differ (Shankar and Mannering, 1998). In addition, there are minute-to-minute traffic flow variations due to interruptions at downstream intersections or bottlenecks. In fact, link travel time variations can be modeled by the type of facility or the type of downstream intersection or bottleneck of the link. In this Chapter, it is assumed that such link travel time variations (or deviations) are an increasing function of the mean travel time of the link. This assumption is consistent with previous empirical studies on the standard deviation of link travel time (Herman and Lam, 1974; Richardson and Taylor, 1978; Taylor 1982). However, it should be noted that there are many possible sources contributing to travel time variations, such as traffic signals, traffic incidents, work zones, weather, etc. Further studies should be carried out to calibrate the relationship between the mean travel time and its variations for different road types under different operating conditions.

In the past decade, much attention has been given to the effects of an unreliable transport network on route choices. The unreliability or uncertainty on the road network may be due to travel time variations as discussed above, degradable roads, travel demand fluctuation, etc. (Asakura and Kashiwadani, 1991; Nicholson and Du, 1997; Bell, 1999). In contrast to deterministic transport networks, stochastic transport networks imply that drivers cannot perfectly predict their travel times. Mirchandani and Soroush (1987) considered that travelers minimize the expected disutilities of uncertain travel times associated with different routes. Uchida and Iida (1993) used the sum of the mean travel time and a safety margin to account for travel time variability. Henn (2000) considered that drivers' route selection is related to their pessimistic/optimistic or risk-taking/risk-averse decisions. Bell (2000) estimated the path choice probability by using the approach of a mixed-strategy Nash equilibrium. Lo and Tung (2003) and Lo et al. (2006) formulated probabilistic models in the form of a travel time budget to account for the effects of travel time reliability due to stochastic degradations in link capacities. Recently, Zhang and Lam (2002) established a reliability-based user equilibrium principle, in which travelers are modeled to select routes so as to maximize the reliability of the minimal path travel time.

Table 8.1 summarizes the principal characteristics and various definitions of road network reliability. Idia and Wakabayashi (1989) proposed a method to determine the terminal reliability of a road network, referred to as the connectivity reliability, by using partial minimal paths and cut sets. Asakura and Kashiwadani (1991) defined some measures of travel time reliability and subsequently modified the conventional traffic assignment model to simulate the fluctuation of road network flow. Asakura (1998)

investigated reliability measures in a road network when some links are possibly damaged by natural disasters. Du and Nicholson (1997) presented an integrated equilibrium model for a large scale, multi-mode degradable transportation system. They further discussed the socio-economic impacts of system degradation, the critical components, and system reliability.

Chen et al. (1999, 2002) introduced capacity reliability as a new network performance index, defined as the probability that the network can accommodate a certain traffic demand at a required service level, while accounting for drivers' route choice behavior. Zhang and Lam (2001) proposed an alternative concept, referred to as travel demand satisfaction ratio, to evaluate the reliability of path travel time. Recently empirical results (Lam and Small, 2001; de Palma and Picard, 2005) indicated that route choices depend not only on the travel times (and costs) of alternative routes but also their reliability. In order to design an efficient and reliable road network, there is a need to develop a traffic assignment model that accounts for both the effects of travel time and reliability. Recently, attention has been given to vulnerability analysis. It is concerned with the identification of links and nodes with significant impacts on network performance. For a road transportation system, vulnerability is defined as the susceptibility of the network to incidents that can result in considerable reductions in road network serviceability, where the serviceability of a link/route/road network is regarded as the possibility to use that link/route/road network at a given time. Berdica (2000, 2002) investigated the effects of a number of input variables, such as demand fluctuations and incidents, on network performance.

In the literature, different approaches have been formulated for the analysis of the reliability of road transportation networks. Generally, they can be characterized by different ways of modeling the effects due to various disturbances and/or uncertainties in travel demand and network capacity, resulting in travel time variations or unreliability. Following a similar vein, in this Chapter, a path preference index (PI) is proposed to quantify the attractiveness of each alternative path in an unreliable transport network. The new feature of the proposed PI is to capture travel time variations due to minute-to-minute traffic flow fluctuations during the same hourly period. Firstly, the path travel time reliability is defined to be a function of the ratio between the path's mean travel time and free-flow travel time. Then by normalization, the path travel time reliability index (RI) and path travel time index (TI) are defined and scaled in the range between [0,100]. The index RI is positively related to the path travel time reliability and TI is negatively related to the path travel time. The weighted sum of RI and TI is referred to as the path preference index (PI). If the PI of a path is equal to 100, this means that travelers on that path travel at its free-flow

travel time with 100% reliability. On the other hand, if the PI of a path approaches 0, its travel time is infinite and travelers have no chance of traveling at its free-flow travel time (i.e. the path travel time reliability approaches 0). Alternatively, it implies that the serviceability or possibility of using this path at free-flow travel time during a given time period is equal to **zero.**

Table **8.1** Principal Characteristics & Various Definitions of Road Network Reliability

Reliability	**Performance Indicator**	**Uncertainty**	**Constraints**	**Probability Definition**
Connectivity (see Iida and Wakayabashi, 1989)	connect or disconnect	Disruption of road links	Not included	connected and disconnected network
Travel Time (I) (see Asakura and Kashiwadani, 1991)	Specified travel time (T)	Fluctuation of daily traffic flow	Constant demand and no link capacity	Travel time less than T
Travel Time (II) (see Asakura, 1998)	Specified network service level (L)	Degradable link capacity	Constant or elastic demand function and link capacity constraint	Service level less than L
System and OD sub-system (see Du and Nicholson, 1997)	Intolerable decrement rate of OD flow (E)	Degradable link capacity	Elastic demand function and link capacity constraint	Decrement rate less than E
Capacity (see Chen et al, 1999, 2002)	Required demand level (Q)	Degradable link capacity	Proportional OD trip table and link capacity constraint	Network reserve capacity greater than Q
Travel time budget (See Lo and Tung, 2003; Lo et al., 2006)	Travel time budget for punctual arrivals	Degradable link capacity	Constant total demand	Within budget time reliability
Travel Demand Satisfaction (Zhang and Lam, 2001)	Specified Travel Demand Satisfaction Ratio (S)	Travel Demand and Degradable link capacity	Elastic demand function	Travel Demand Satisfaction Ratio less than S
Road Vulnerability (Berdica, 2000, 2002)	Travel time and serviceability	Demand fluctuations and incidents	Consequence of incidents	Probability for an incident to occur

For the long-term network design problems, this Chapter presents a reliability-based user equilibrium (RUE) model in which both path travel time and reliability are considered and accounted for by the normalized path preference index PI. It can be used to assess the impacts of the values of travel time and reliability on drivers' route choice behavior, traffic flow patterns, and system performance (i.e. network travel time and reliability). By setting the weights on the two indices, RI and TI, through a linear combination that sums to one, then one can allocate different importance to each of them. In the extreme cases, if the weight set on TI (RI) is 1.0

(0.0), the traffic assignment results of the proposed RUE model are equivalent to the conventional user equilibrium (UE) model. On the other hand, if the weight set on RI (TI) is 1.0 (0.0), the travel time reliability on all the used paths by OD pair is equal to each other. Hence, the RUE model contains the conventional UE principle as one special case.

This Chapter is organized as follows. In the next section, the path preference index is defined and discussed together with the definitions of the path travel time and reliability indices. The mathematical formulations and proofs of the RUE are then given. The example network with Braess paradox is used to illustrate the characteristics of the proposed RUE model. Finally, conclusions are drawn together with some recommendations for further studies.

8.2 Path Preference and Related Indexes

The path preference index is defined as the weighted sum of the path travel time reliability index and path travel time index on a particular path. In the following, after defining the path travel time reliability and path travel time indices, the path preference index is presented.

The proposed RUE model is formulated for the purpose of long-term transport planning, while capturing variations in link travel times due to minute-to-minute traffic flow fluctuations in the road network. Let c_j^{rs} be the travel time on path j from origin r to destination s, expressed as:

$$c_j^{rs} = \pi_j^{rs} + \varepsilon_j^{rs}, \quad \forall r, s, j \tag{1}$$

where π_j^{rs} is the mean travel time on path j from origin r to destination s and ε_j^{rs} is the random term on path j with $E[\varepsilon_j^{rs}] = 0$ resulting from temporal fluctuations in traffic flow (Lam and Small, 2001).

8.2.1 The Path Travel Time Reliability Index

Previous studies have paid insufficient attention on travel time reliability (Bates et al., 2001). Virtually all previous studies worked with hypothetical scenarios, partly due to the fact that measuring the variability of travel times with actual data is difficult and time consuming. Asakura (1998) proposed a performance measure for the travel time reliability based on the ratio between the mean travel time and free-flow travel time, which is eas-

ily understood by drivers. The mean travel time π_j^{rs} is known as the resultant actual average travel time on path j from origin r to destination s. Denote π_f^{rs} as the free-flow travel time from origin r to destination s. Therefore, the performance measure proposed in this Chapter is $\pi_j^{rs} / \pi_f^{rs} (\forall r, s, j)$.

This scaled free-flow travel time measure can be adopted as the benchmark for assessing the level of service (LOS) on alternative paths (Lam and Zhang, 2001). By expressing π_j^{rs} as $\theta_R \pi_f^{rs}$, the parameter θ_R, a predefined performance measure, can be used to indicate the LOS on a particular path j. The Highway Capacity Manual (National Research Council, 2000) introduces the concept of "level of service" as a qualitative measure of the composite effects of operating speed and travel time on the highway facilities. In practice, average operating speed and free-flow speed are used to categorize the level of service. Six levels are established, A to F, where LOS A and F are referred to as the free-flow (non-congested) and forced-flow (very congested) conditions, respectively. On the basis of this "level of service" concept, the LOS proposed in this Chapter can be expressed in terms of θ_R with reference to the ratio of the free-flow speed and average operating speed (μ). For instance, as shown in Exhibit 15-2 of the Highway Capacity Manual (National Research Council, 2000), the LOS criteria for urban road with average free-flow speed of 65 kilometers per hour (km/hr) can be defined as below:

LOS	Description	μ (km/hr)	θ_R
A	Free flow, high speed	> 59	< 1.1
B	Stable flow, reasonable freedom	> 46	< 1.4
C	Stable flow, higher volume	> 33	< 2.0
D	Unstable flow, tolerable speed	> 26	< 2.5
E	Unstable flow, at-capacity volume	> 21	< 3.1
F	Forced flow, spillback queues	$\leq$ 21	$\geq$ 3.1

The random variable of path travel time is considered to follow the Normal distribution. Denote x as the standard Normal variate associated with the path travel time and σ_j^{rs} as the standard deviation of travel time. From empirical studies, σ_j^{rs} is found to be a function of the mean travel time π_j^{rs} (Richardson and Taylor, 1978; Taylor 1982; Shankar and Mannering, 1998), i.e., $\sigma_j^{rs} = \sigma_j^{rs}(\pi_j^{rs})$. The path travel time reliability is defined as the probability of the path travel time c_j^{rs} being less than the

specified threshold $\pi_f^{rs}\theta_R$ associated with a particular LOS. Mathematically, it can be expressed as:

$$R_j^{rs}(\omega_j^{rs}) = P\left(c_j^{rs} \le \pi_f^{rs}\theta_R\right) = P\left(x \le \pi_f^{rs}(\theta_R - \omega_j^{rs})/\sigma_j^{rs}(\omega_j^{rs})\right), \ \forall r,s,j \quad (2)$$

where $\omega_j^{rs} = \pi_j^{rs}/\pi_f^{rs}$ is the performance measure and $x = (c_j^{rs} - \pi_j^{rs})/\sigma_j^{rs}$ transforms c_j^{rs} to the standard normal variate x. σ_j^{rs} can be rewritten as the form of $\sigma_j^{rs} = \sigma_j^{rs}(\omega_j^{rs})$. Eqn. (2) satisfies $\frac{dR_j^{rs}}{d\omega_j^{rs}} < 0$, implying that the path travel time reliability is a monotonic decreasing function of the performance measure ω_j^{rs} or the mean travel time. The path travel time reliability is decreasing when the mean travel time increases.

To normalize this measure, the path travel time reliability index (RI) is scaled from 0 to 100 and defined by rewriting (2) as:

$$RI_j^{rs}(\omega_j^{rs}) = 100R_j^{rs}(\omega_j^{rs}) \ \text{ and } \ \frac{dRI_j^{rs}}{d\omega_j^{rs}} < 0, \ \forall r,s,j \quad (3)$$

The path travel time reliability index is also a monotonic decreasing function of the performance measure ω_j^{rs}. With θ_R set to be one (i.e. LOS A), the path travel time reliability index approaches 100 implies that travelers on this path can travel at free-flow travel time with 100% reliability. On the other hand, when the path travel time reliability index is 0, travelers have no chance of arriving at their destination at free-flow travel time by using this path. In other words, it can be interpreted that the serviceability or possibility of using this path at LOS A during a given time period is equal to zero.

8.2.2 The Path Travel Time Index

The path travel time index (TI) is defined mathematically as:

$$TI_j^{rs}(\omega_j^{rs}) = 100e^{-\theta_T(\omega_j^{rs}-1)} \ \text{ and } \ \frac{dTI_j^{rs}}{d\omega_j^{rs}} < 0, \ \forall r,s,j \quad (4)$$

where θ_T is a scaled parameter. The value of θ_T can be calibrated and considered as a scaled value of time for driver route choices. The path travel time index is a monotonic decreasing function of the performance measure ω_j^{rs}. If the mean travel time is equal to the free-flow travel time or the performance measure ω_j^{rs} equals 1.0, the path travel time index is $TI_j^{rs}(1.0) = 100$. If the mean travel time approaches infinity, on the other hand, the path travel time index becomes 0.

8.2.3 The Path Preference Index

By integrating both the path travel time and reliability indices for ranking drivers' preferences to alternative routes between each OD pair, the path preference index (PI) is defined mathematically as:

$$PI_j^{rs}(\omega_j^{rs}) = \alpha RI_j^{rs}(\omega_j^{rs}) + (1-\alpha)TI_j^{rs}(\omega_j^{rs}) \quad \text{and} \frac{dPI_j^{rs}}{d\omega_j^{rs}} < 0 \ \forall r,s,j \qquad (5)$$

where $1 \geq \alpha \geq 0$ is a weight allocated to the path travel time reliability index and $100 \geq PI_j^{rs}(\omega_j^{rs}) \geq 0$. If $\alpha = 0$, only the path travel time index is considered by drivers for their route choices. If $\alpha = 1$, only the path travel time reliability index is taken into account. Eqn. (5) implies that the destination can (cannot) be accessed at free-flow travel time (i.e. LOS A) on the chosen path when the path preference index approaches to 100 (0). The PI is a monotonic decreasing function of the performance measure ω_j^{rs} and is proposed as an indicator for ranking drivers' preferences to alternative routes. In addition, this indicator can be extended to be a network performance indicator, as to be discussed later.

8.3 The RUE Formulation Based on the Path Preference Index

The road network is formulated by a directed graph G(N, A), where N is a set of nodes and A is a set of directed links. The set of nodes includes subsets of centroids and intersections. A path (or route) is a set of sequential links and nodes without loop from an origin (a centroid) to a destination

(another centroid). Let v_a be the traffic flow on link a. The link travel time is

$$t_a(v_a) = \bar{t}_a(v_a) + \tau_a, \ \forall\ a \tag{6}$$

where $\bar{t}_a(v_a)$ is the mean travel time on link a; τ_a is a random travel time term incurred by link flow variations with $E[\tau_a] = 0$; $\bar{t}_a(v_a)$ is a strictly increasing function of the link flow v_a, $\frac{d\bar{t}_a}{dv_a} > 0$ (if $v_a > 0$).

The travel time c_j^{rs} on path j is

$$c_j^{rs} = \sum_a \delta_{aj}^{rs} t_a^{rs}(v_a), \forall r, s, j \tag{7}$$

where $\delta_{a,j}^{rs}$ is the link-path incidence matrix, $\delta_{a,j}^{rs}$=1 if link a is on path j, and 0 otherwise.

The mean travel time on path j is

$$\pi_j^{rs} = E[c_j^{rs}] = \sum_a \delta_{aj}^{rs} E[t_a(v_a)] = \sum_a \delta_{aj}^{rs} \bar{t}_a(v_a), \forall r, s, j. \tag{8}$$

If the travel time on link a is independent from that of other links in the road network, the variance of the path travel time on route j via link a is then

$$\sigma_{rsj}^2 = Var[c_j^{rs}] = \sum_a \delta_{aj}^{rs} Var[t_a(v_a)] = \sum_a \delta_{aj}^{rs} \sigma_a^2, \forall r, s, j. \tag{9}$$

where σ_a^2 is the variance of travel time on link a. Denote $\bar{t}_{af}$ as the free-flow travel time on link a. Then the free-flow travel time from origin r to destination s is:

$$\pi_f^{rs} = \min_j \sum_a \delta_{aj}^{rs} \bar{t}_{af}. \tag{10}$$

Let f_j^{rs} be the traffic flow on path j and PI^{rs} be the maximal path preference index from origin r to destination s. The nonlinear complementary problem (NCP) for the RUE principle based on the proposed PI is formulated mathematically as follows:

$$f_j^{rs}(PI_j^{rs} - PI^{rs}) = 0, \forall r, s, j \tag{11}$$

$$PI_j^{rs} - PI^{rs} \le 0, \forall r, s, j \tag{12}$$

$$\sum_j f_j^{rs} = q^{rs}, \forall r, s, \tag{13}$$

$$\sum_{r,s}\sum_{j}\delta_{aj}^{rs}f_j^{rs} = v_a, \forall a \tag{14}$$

$$f_j^{rs} \geq 0, \forall r, s, j. \tag{15}$$

Note that eqns. (11) and (12) are the complementary slackness conditions: $PI_j^{rs} - PI^{rs} < 0$, $f_j^{rs} = 0$ and $PI_j^{rs} - PI^{rs} = 0$, $f_j^{rs} \geq 0$. As a result, when the path preference index on route j is smaller than the maximum PI^{rs}, the path flow on route j is equal to zero. If the path preference index on route j is equal to the maximum PI^{rs}, then the path flow on route j is equal to or greater than zero. These RUE conditions are equivalent to those for the user equilibrium principle (Wardrop, 1952) mathematically. Eqn. (13) is the path flow constraint with respect to the OD demand while eqn. (14) states that summing the path flows for all OD pairs via link a should be equal to the traffic flow on that link. Eqn. (15) is the non-negative constraint for the path flow.

This nonlinear complementary problem can be easily transformed to an equivalent variational inequality (VI) problem. Let $\mathbf{f}$ be the vector of path flows $\{f_j^{rs}\}$, and $\mathbf{F}(\mathbf{f})$ be the vector of $\{PI^{ij} - PI_k^{ij}\}$.

Proposition 1 The NCP (11)-(15) is equivalent to the VI problem below:

Find $\mathbf{f}^* \in \Omega$ such that: $(\mathbf{f} - \mathbf{f}^*)^T \mathbf{F}(\mathbf{f}^*) \geq 0 \quad \forall \mathbf{f} \in \Omega$ (16)

where Ω is the constraint set of path flows

$$\Omega = \left\{ f_j^{rs} \geq 0, \forall r, s, j, \sum_j f_j^{rs} = q^{rs}, \forall r, s \right\}.$$

Proof See proposition 1.4 in Nagurney (1999).

Theorem 1 The VI problem (16) admits at least one solution $\mathbf{f}^* \in \Omega$.

Proof

The VI formulation (16) can be stated as

$$\text{Find } \mathbf{f}^* \in \Omega \text{ such that } \sum_{rs}\sum_{j}(f_j^{rs} - f_j^{rs^*})(PI^{rs^*} - PI_j^{rs^*}) \geq 0 \quad \forall \mathbf{f} \in \Omega \tag{17}$$

$$\Leftrightarrow \sum_{rs}\sum_{j}(f_j^{rs} - f_j^{rs^*})PI^{rs^*} - \sum_{rs}\sum_{j}(f_j^{rs} - f_j^{rs^*})PI_j^{rs^*} \geq 0 \quad \forall \mathbf{f} \in \Omega$$

$$\Leftrightarrow \sum_{rs} PI^{rs^*}\left(\sum_{j} f_j^{rs} - f_j^{rs^*}\right) - \sum_{rs}\sum_{j}(f_j^{rs} - f_j^{rs^*})PI_j^{rs^*} \geq 0 \quad \forall \mathbf{f} \in \Omega \tag{18}$$

It follows from $\mathbf{f} \in \Omega, \mathbf{f}^* \in \Omega$

$$\sum_j f_j^{rs} = q^{rs}$$

$$\sum_j f_j^{rs^*} = q^{rs}$$

Therefore, the first term of the left-hand in eqn. (18) is zero

$$\sum_{rs} PI^{rs^*} \left(\sum_j f_j^{rs} - f_j^{rs^*}\right) = 0 \tag{19}$$

It follows from eqns. (18) and (19) that the VI formulation (16) is equivalent to VI below

$$\text{Find } \mathbf{f}^* \in \Omega \text{ such that } \sum_{rs}\sum_j (f_j^{rs} - f_j^{rs^*})(-PI_j^{rs^*}) \geq 0 \quad \forall \mathbf{f} \in \Omega \tag{20}$$

Suppose that $\mathbf{G}(\mathbf{f})$ is the vector of $\left\{-PI_j^{rs^*}\right\}$, eqn. (20) can then be stated as

$$\text{Find } \mathbf{f}^* \in \Omega \text{ such that } \left(\mathbf{f} - \mathbf{f}^*\right)^T \mathbf{G}\left(\mathbf{f}^*\right) \geq 0 \quad \forall \mathbf{f} \in \Omega \tag{21}$$

Obviously, the constraint set Ω is a compact convex and $\mathbf{G}(\mathbf{f})$ is continuous on Ω. According to the theorem 1.4 (Nagurney 1999), the existence of a RUE solution is verified.

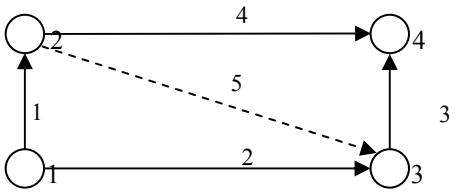
It should be pointed out that the RUE model is based on the assumption that the route set is given and fixed for each OD pair within the road transportation network. According to the empirical study by Cascetta et al.(1996), the number of actually chosen routes between an OD pair is generally in the range of 6 to 8 even for large-scale road networks. Therefore, the route set between each O-D pair can be pre-determined using the actually chosen route obtained by observation and/or interview surveys.

8.4 Numerical Example

The purposes of the numerical example are to illustrate: (i) the differences on traffic assignment results between the proposed RUE and conventional UE models; (ii) the effects of the RUE principle on the well-known Braess paradox (Lam, 1988) when drivers do consider travel time reliability for their route choices; (iii) the impacts of different weights (α) on the travel time reliability with various OD demands; and (iv) the resultant TI, RI, and PI when α= 0.38 (i.e. 11.90/(19.22+11.90)=0.38) – the α value based on

empirical findings for route choice models (Lam and Small, 2001) considering the value of time to be $19.22 per hour and the value of reliability to be $11.90 per hour.

The example network, shown in Figure 8.1 with the Braess paradox (Lam, 1988), is used in this Chapter to facilitate the presentation of the essential ideas and to illustrate the performance of the proposed RUE model. The example network consists of 5 links, 1 OD pair (nodes: 1→4) and 3 paths (links: 1→4, 2→3 and 1→5→3).



Path 1: Links 1, 4
Path 2: Links 2, 3
Path 3: Links 1, 5, 3

Fig. 8.1 The Example network

In the Braess paradox, link 5 is the proposed additional link that would lead to increases in the total network total time in the example network. For this numerical example, a link travel time function similar to the one adopted by the U.S. Bureau of Public Roads (BPR) is used:

$$\bar{t}_a = \bar{t}_{af} + B_a (v_a)^4 \quad \text{(in minutes)}$$

where $\bar{t}_{af}$ is the free-flow travel time on link a; B_a is the congestion or delay coefficient on link a, with the link flow expressed in thousand vehicles per hour (veh/h).

On Interstate 90 (I-90) located some 50 kilometers east of Seattle, Shankar and Mannering (1998) conducted surveys to calibrate the relationship between speed-deviation and speed as follows:

$$\text{Speed-deviation}=\text{Constant}_1-\text{Constant}_2\times\ln(\text{Mean-speed})+\text{Other-Factors}$$

where $\text{Constant}_1 \geq 0$ and $\text{Constant}_2 \geq 0$. The other factors are related to the environmental and seasonal factors and are captured as part of Constant_1 in this Chapter for simplicity.

As Mean-travel-time = Distance/(Mean-speed), we have: ln (Mean-travel-time) = ln (Distance) -ln (Mean-speed). In this example, we focus on the impact of the mean travel time on the standard deviation of link travel time and hence rewrite the standard deviation formula as:

$$\sigma_a = \beta_{a0} + \beta_{a1} \ln \bar{t}_a , \tag{22}$$

where $\beta_{a0} \geq 0$ and $\beta_{a1} \geq 0$ are coefficients related to link a. β_{a0} is the standard deviation when the mean travel time is equal to 1.0 unit. $\beta_{a1} / \bar{t}_a$ is the increasing rate of the standard deviation when the mean travel time is increased by 1.0 unit. Similar to the BPR function for link travel time, these coefficients β_{a0} and β_{a1} should be calibrated by road type in accordance with the sources causing the link flow variations.

Table 8.1 shows the input link data of the example network. It can be seen that the coefficients of links 1 and 3 are exactly the same; similarly, the coefficients of links 2 and 4 are also equal to each other. Due to the symmetry of the example network, the traffic assignment results on path 1 (i.e. links: 1→4) are similar to that on path 2 (i.e. links: 2→3).

From eqn. (10) and Table 8.1, the referred free-flow travel time on path 3 is π_f^{rs}= 40+40+15.4=95.4 (min). On the basis of eqns. (3) to (5), the scaled parameters are defined/calibrated to be: $\theta_R = 2.0$ (i.e. LOS D) and $\theta_T = 0.6931$, so that the path travel time reliability index $RI_3(2.0) = 50.00$ and the path travel time index $TI_3(2.0) = 50.00$ in order to maintain the symmetric characteristics of the example network.

Table 8.1 Input Data of the Example network

Link Number	$\bar{t}_{af}$ (min)	B_a	β_{a0} (min)	β_{a1}
1	40	0.5	30.0	2.0
2	185	0.9	140.0	1.0
3	40	0.5	30.0	2.0
4	185	0.9	140.0	1.0
5	15.4	1.0	25.0	3.0

With these input data and different weights (α) on the travel time reliability, the proposed RUE model is used to estimate the link travel times and link flows together with path flows on this example network. The RUE traffic assignment results are given in Table 8.2 and compared with those of the conventional UE model.

It can be observed in Table 8.2 that both the travel time and reliability have significant effects on the traffic assignment results. For instance, the link flow and link travel time are 2000 (veh/h) and 31.4 (min) on the additional link (i.e. link 5) under the UE condition wherein the link travel time reliability is 49.32% at α =0.0. When drivers do consider both travel time

and reliability for their path choices, with α =0.38, the link flow, link travel time, and link travel time reliability change to 1536 (veh/h), 21 (min), and 61.30%, respectively. We note that the link travel time reliability can be improved to 66.22% while the link flow and travel time can be reduced to 1080 (veh/h) and 16.8 (min), respectively, if route choices are dependent on reliability only (i.e. α=1.0). As a result, the total network travel time reliability is enhanced from 48.08% to 71.12% and further to 81.59%. The total network travel time is decreased from 2204.4 to 2018.7 and further to 1929.1 (thousand veh-min). This result illustrates that the RUE results vary with different values of α (i.e. the weight allocated on the travel time reliability for driver route choice).

Table 8.2 The Equilibrium Travel Time (min), Flow (veh/h) and Travel Time Reliability (TTR) (%) by Link, Path and Network (OD demand =6000 veh/h)

Item	α = 1.0 on the RI			α = 0.38 on the RI			α =0.0 on the RI			UE	
	Flow	Time	TTR	Flow	Time	TTR	Flow	Time	TTR	**Flow**	**Time**
Link 1	3540	118.5	**16.52**[a]	3768	140.8	6.38	4000	168.0	1.44	**4000**	**168.0**
Link 2	2460	217.9	85.23	2232	207.3	86.85	2000	199.4	87.98	**2000**	**199.4**
Link 3	3540	118.5	16.52	3768	140.8	6.38	4000	168.0	1.44	**4000**	**168.0**
Link 4	2460	217.9	85.23	2232	207.3	86.85	2000	199.4	87.98	**2000**	**199.4**
Link 5	**1080**	**16.8**	**66.22**	**1536**	**21**	**61.30**	**2000**	**31.4**	**49.32**	**2000**	**31.4**
Path 1	2460	336.4	**77.46**[b] **(16.69**[c]**)**	2232	348.1	**75.05 (14.93)**	2000	367.4	70.81 (12.15)	**2000**	**367.4**
Path 2	2460	336.4	**77.46 (16.69)**	2232	348.1	**75.05 (14.93)**	2000	367.4	70.81 (12.15)	**2000**	**367.4**
Path 3	1080	253.8	**16.69 (16.69)**	1536	302.6	**4.50 (4.62)**	2000	367.4	0.42 (0.42)	**2000**	**367.4**
Net-work	6000[e]	**1929.1**[f]	**81.59**[d]	6000	**2018.7**	**71.12**	6000	**2204.4**	**48.08**	**6000**	**2204.4**

Note: [a] Link Travel Time Reliability (%) = $P\left(x \le (\theta_R \bar{t}_{af} - \bar{t}_a)/\sigma_a\right)$x100%

[b] Path Travel Time Reliability (%) = $P\left(x \le (\theta_R \pi_{fj}^{rs} - \pi_j^{rs})/\sigma_{rsj}\right)$ x100% where π_{fj}^{rs} is the free-flow travel time on path j.

[c]Travel Time Reliability (%) = $P\left(x \le (\theta_R \pi_f^{rs} - \pi_j^{rs})/\sigma_{rsj}\right)$ x100%

[d] Network Travel Time Reliability (%) = $P\left(x \le (\theta_R \sum_a \bar{t}_{af} v_a - \sum_a \bar{t}_a v_a)/(\sum_a \sigma_a^2 v_a^2)^{1/2}\right)$ x100%

[e]Total OD Demand (veh/h) in the network

[f]The Total Network Travel Time = $\sum_a \bar{t}_a v_a$ (thousand veh-min)

In addition, it should be noted that θ_R is set to be 2.0 (i.e. LOS D) in this numerical example. When α =0.38, the travel time reliability was found to be 61.30%, 75.05% and 71.12% for link 5, path 1 and the whole network respectively. It implies that the serviceability or probability of using this link/path/network at LOS D during the study period is equal to 61.30%, 75.05% and 71.12% respectively. Obviously, we can reset the value of θ_R for different LOS and then re-calculate the serviceability of a link/path/road network by LOS at a given time period. Following the same vein of Berdica (2000, 2002) regarding the vulnerability and serviceability analysis, the proposed reliability measures can be used for identification of links and paths with significant impacts on network performance due to various incidents and/or road expansion projects.

It can be seen in Table 8.2 that the link, path, and network travel time reliabilities all increase with α. For higher α values, more traffic is allocated onto the less congested and more reliable routes, which, however, may have longer travel distances and times. In a sense, there is a tradeoff between reliability and travel time.

Table 8.3 shows the effects of α and OD demand on the total network travel time. The results show that the total network travel times decrease with reduction in OD demand but increase with α. Except for the case when the OD demand is sufficiently low, equal to 2000 in this example, the total network travel time is unchanged for different values of α.

We can assess the impacts of the RUE principle on the Braess paradox by investigating the results in Tables 8.2 and 8.3. It can be seen that the total network travel time and the network travel time reliability in the example network are improved if drivers select routes based on the RUE principle. At the OD demand of 6000 (veh/h), the Braess paradox occurs in the UE assignment results whereas the total network travel time increases from 2030.4 to 2204.4 (thousand veh-min) when link 5 is added to the example network. However, the total network travel time can be reduced and lowered than 2030.4 (thousand veh-min) in the RUE assignment when α = 0.38. In this example network, the Braess paradox disappears in the RUE results when the weight α set on the travel time reliability is greater than a certain value. In this example network when α = 0.38, the total network travel time decreases to 2018.7 (thousand veh-min) and the network travel time reliability increases to 71.12%. In the extreme case when driver route choice relies wholly on the travel time reliability with α = 1.0, the total network travel time is 1929.1 (thousand veh-min). It approaches to the system optimal result of 1914.9 (thousand veh-min) as obtained by Lam (1988). It is because more traffic is shifted to the less congested and more reliable routes whereas the network travel time reliability is enhanced to 81.59%.

Based on this preliminary result, it is too early to say whether by doing so will always avoid the Braess paradox. But certainly, adding another criterion (in this case, reliability) to capture travelers' route choices would alter the traffic assignment pattern. In particular, the reliability criterion tends to spread some drivers away from the shortest and at the same time congested routes, thus allowing for some sort of balancing of route flows in the network, resulting in a reduction in the overall network travel time.

Table 8.3 The Total Network Travel Time (10^3veh-min) at Equilibrium

OD Demand (veh/h)	2000	3000	4000	5000	**6000**
α = 1.00 on the RI	254.8	671.1	993.2	1378.4	**1929.1**
α = 0.38 on the RI	254.8	721.3	1058.2	1462.1	**2018.7**
α = 0.00 on the RI	254.8	772.2	1169.5	1617.6	**2204.4**
Before adding link 5	452.8	696.3	989.6	1398.4	**2030.4**

Table 8.4 shows that both the path travel time index (TI) and the path travel time reliability index (RI) decrease in the example network with and without link 5 as the OD demand increases and α is fixed at 0.38. It can be seen that for the case with link 5 and the OD demand is 2000, both the TI and RI of path 3 are greater than those of the alternative paths as all the trips are allocated to path 3. For the other cases, in general, if the path travel time index (TI) is higher, then the path travel time reliability index (RI) is lower than that of the alternative path. When the OD demand is 4000 and link 5 is added, the TI is 27.01 on path 1, which is less than the TI of 39.00 on path 3, and the RI of 27.08 on path 1 is greater than the RI of 17.30 on path 3. For the network without 5, the TIs are both 20.33 on paths 1 and 2 while the RIs are both 33.92 on paths 1 and 2. These results illustrate drivers' trade-offs between the path travel time and the path travel time reliability when making their route choices.

The effects of the RUE principle on driver route choices in terms of the path preference index (PI) are illustrated in Table 8.5 and Figure 8.2. It can be observed in Table 8.5 that when the OD demand is 2000 in the example network with link 5 and α = 0.38, the PI on path 3 is 81.04, which is greater than those on paths 1 and 2. As a result, all the trips are assigned to path 3 only. On the other hand, when the OD demands are greater than 2000, the OD trips are distributed among all the alternative paths, with all paths attaining the same PI. It should be noted that the PI decreases when the path travel time increases. Similar to the example network without link 5, the PIs on path 1 decrease when the path travel time increases. Thus, the path preference index is a robust indicator to reflect both the level of con-

gestion and the degree of reliability on a particular path for travel between an OD pair.

Table 8.4 The Equilibrium RI, TI v.s. OD Demand with $\alpha = 0.38$

OD Demand (veh/h)			2000	3000	**4000**	5000	6000
With Link 5	Path Travel Time Index (TI)	Path 1	36.80	29.70	**27.01**	21.40	15.94
		Path 3	79.26	35.03	**39.00**	27.89	22.20
	Path Travel Time Reliability Index (RI)	Path 1	39.04	31.78	**27.08**	21.94	14.93
		Path 3	83.96	23.28	**17.30**	11.45	4.62
Without Link 5	The TI on Path 1		25.33	23.86	**20.33**	14.49	7.83
	The TI on Path 2		25.33	23.86	**20.33**	14.49	7.83
	The RI on Path 1		39.72	38.13	**33.92**	25.73	13.97
	The RI on Path 2		39.72	38.13	**33.92**	25.73	13.97

Table 8.5 The User Equilibrium PI v.s. Travel Demand with $\alpha = 0.38$

	OD Demand (veh/h)		**2000**	3000	4000	5000	6000
With Link 5	$\omega_1^{rs} = \pi_1^{rs} / 95.4$		**2.44**	2.75	2.97	3.22	3.65
	$\omega_3^{rs} = \pi_3^{rs} / 95.4$		**1.34**	2.51	2.61	2.84	3.17
	Path Preference Index (PI)	Path 1	**37.65**	30.49	26.15	21.64	15.57
		Path 3	**81.04**	30.49	26.15	21.64	15.57
Without Link 5	$\omega_1^{rs} = \pi_1^{rs} / 95.4$		**2.37**	2.43	2.59	2.93	3.55
	The PI on Path 1		**34.25**	32.71	28.76	21.46	11.60
	The PI on Path 2		**34.25**	32.71	28.76	21.46	11.60

Figure 8.2 shows the variations of PI on the chosen paths in the example networks with and without link 5. The PI in network with link 5 is generally greater than that without link 5, implying that the PI is improved with the addition of link 5. In summary, the Braess paradox may disappear in the example network with the new link when drivers consider both travel time and reliability for their route choices (Lam and Small, 2001). It was found in the numerical example that the Braess paradox is obviated when the weight α on the travel time reliability is set equal to or greater than 0.38. However, it must be noted that the proposed RUE model does not preclude the occurrence of this paradox, particularly when UE is a special case of RUE when α is set to be zero.

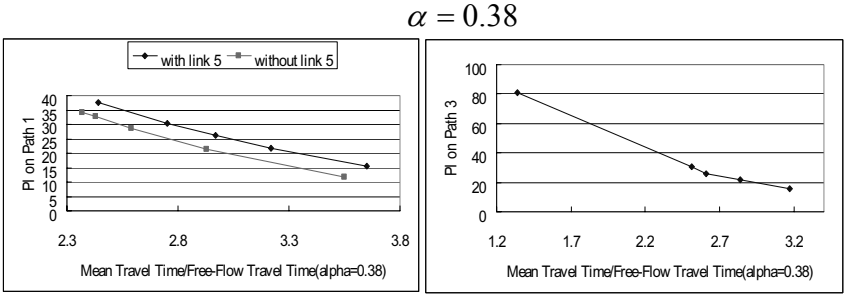


Fig. 8.2 The RUE Results: Path Preference Index (PI) v.s. $\omega_j^{rs} = \pi_j^{rs} / \pi_f^{rs}$

8.5 Conclusions

This Chapter proposed a new reliability-based user equilibrium (RUE) model by taking into account the effects of both travel time and reliability on driver route choice behavior. We considered network uncertainty in terms of link travel time variations due to temporal fluctuations of traffic flows within the hourly period. The results revealed that both travel time and reliability have significant impacts on driver route choices, the resultant traffic flow pattern, and network performances. The proposed reliability measures can be used for identification of links and paths with significant impacts on network performance due to incidents and/or road expansion projects, which can be part of the vulnerability and serviceability analysis for a network.

It is well-known that link addition to a network might introduce the Braess paradox, possibly leading to degradations of the total network travel time under the UE principle. However, under the proposed RUE principle, such paradox might be avoided when drivers consider both travel time and reliability on their route choices. Adding another criterion (in this case, reliability) to capture travelers' route choices would alter the traffic assignment pattern. In particular, the reliability criterion tends to spread some drivers away from the shortest and at the same time congested routes, thus allowing for some sort of balancing of route flows in the network. The end result is to benefit the overall network travel time. However, the Braess paradox may not be completely obviated particularly when the weight α on the travel time reliability is close to zero, i.e., when drivers pay very little attention to the path travel time reliability for their route choices.

The numerical example also showed that the UE solution is a special case of the RUE when the weights are set to be 1.0 on the index of travel time or 0.0 on the index of reliability. In this Chapter, the relationship of the

path travel time and reliability was investigated and formulated as the path preference index (PI) in order to rank the driver preference to alternative routes. On the basis of the PI, a robust RUE model was formulated mathematically. The existence of the RUE solution was also established theoretically.

Some future extensions of this research work remain. The relationship between mean travel time and travel time standard deviation should be calibrated by link type with empirical data. The network design problem on the basis of the RUE principle should be further studied, particularly under non-recurrent congestion due to incidents. The proposed model can also be extended to road networks with multi-user classes and elastic demand. An efficient solution algorithm needs to be developed to determine the route set for the RUE model.

Acknowledgements

The work described in this Chapter was supported by a joint research grant from the Research Grant Council of the Hong Kong Special Administration Region to the Hong Kong Polytechnic University (Project No. N_PolyU 515/01) and the National Natural Science Foundation of China (Project No. 70131160744) to the BeiHang University. This research was also sponsored by the Competitive Earmarked Research Grant HKUST6033/01E from the Hong Kong Research Grant Council and the internal research grant PolyU 1-ZE10 from the Hong Kong Polytechnic University.

References

Asakura, Y. and M. Kashiwadani. 1991. Road Network Reliability Caused by Daily Fluctuation of Traffic Flow. *Proceedings of the 19th PTRC Summer Annual Meeting*, Brighton, England, Seminar G, 73-84.

Asakura, Y. 1998. Reliability Measures of an Origin and Destination Pair in a Deteriorated Road Network with Variable Flow. In: *Transportation Networks: Recent Methodological Advances*. (Ed. M.G.H. Bell), Pergamon Press: Oxford: 273-287.

Berdica, K. 2000. *Analysing Vulnerability in the Road Transportation System.* KUNGL. TEKNISKA HŐGSKOLAN Royal Institute of Technology, Department of Infrastructure Planning, Division of Transport and Location Analysis, Sweden. Report Reference No. TRITA-IP FR 00-76. ISSN 1104-683X.

Berdica, K. 2002. An Introduction to Road Vulnerability: What has been done, is done and should be done. *Transport Policy,* (9)2, 117-127.

Bates, John, Polak John, Peter Jones, and Cook Andrew. 2001. The Valuation of Reliability for Personal Travel. *Transportation Research*, 37E, 191-229.

Bell, M.G.H. 1999. Measuring Network Reliability: A Game Theoretic Approach. *Journal of Advanced Transportation*, 33(2), 135-46.

Bell, M.G.H. 2000. A Game Theory Approach to Measuring the Performance Reliability of Transport Networks. *Transportation Research*, 34B, 533-545.

Cascetta, E. 2001. *Transportation Systems Engineering: Theory and Methods*. Kluwer Academic Publisher, Dordrecht, The Netherland.

Cascetta, E., A. Nuzzolo, F. Russo and A. Vitetta. 1996. A Modified Logit Route Choice Model Overcoming Path Overlapping Problems. Specification and Some Calibration Results for Interurban Networks. *Proceedings of the 13th International Symposium on the Theory of Road Traffic Flow*, 697-711.

Chen, A, H. Yang, H.K. Lo, and W.H. Tang. 1999. A Capacity Related Reliability for Transportation Network. *Journal of Advanced Transportation*, 33(2), 183-200.

Chen, A, H. Yang, H, Lo, H.K. and W.H. Tang. 2002. A Capacity Reliability of a Road Network: An Assessment Methodology and Numerical Results. *Transportation Research*, 36B, 225-252.

Dafermos S.C. and F.T. Sparrow. 1968. The Traffic Assignment Problem for A General Network. *National Bureau of Standards J. Res.*, 73B, 91-118.

Dagazo, C.F. and Y. Sheffi. 1977. On Stochastic Models of Traffic Assignment. *Transportation Science*, 11(3), 253-274.

de Palma, A., and N. Picard. 2005. Route Choice Decision under Travel Time Uncertainty. *Transportation Research* 39A, 295-324.

Henn, V. 2000. Fuzzy Route Choice Model for Traffic Assignment. *Fuzzy Sets and Systems*, 116 (1), 77-101.

Herman, R.E. and Lam, T. 1974. Trip Characteristics of Journeys to and from Work. In Buckley, D.J. (Ed.), *Proceedings of the 6th International Symposium on Transportation and Traffic Theor*y. Elsevier, New York, 57-85.

Idia, Y. and H. Wakabayashi. 1989. An Approximation Method of Terminal Reliability of A Road Network Using Partial Minimal Path and Cut Set. *Proceedings of the 5th WCTS*, Yokohama, 367-380.

Lam, W.H.K. 1988. Effects of Road Pricing on System Performance. *Traffic Engineering and Control*, 29, 631-635.

Lam, T.C. and K.A. Small. 2001. The value of Time and Reliability: Measurement from A Value of Pricing Experiment. *Transportation Research*, 37E, 231-251.

Lo, H.K. and Y.K. Tung. 2003. Network with Degradable Links: Capacity Analysis and Design. *Transportation Research*, 37B, 345-363.

Lo, H.K., X.W. Luo, and B. Siu. 2006. Degradable Transport Network: Travel Time Budget of Travelers with Heterogeneous Risk Aversion. *Transportation Research B*. In press.

Mirchandani, P. and H. Soroush. 1987. Reliabilistic Traffic Equilibrium with Probabilistic Travel Times and Perceptions. *Transportation Science*, 21(3), 133-152.

Nagurney, A. 1999. *Network Economics: A Variational Inequality Approach*. Kluwer Academic Publishers, Boston, Massachusetts.

National Research Council (2000) *Highway Capacity Manual*. Washington, D.C., U.S.A.

Nicholson, A.J. and Z.P. Du. 1997. Degradable Transportation Systems: An Integrated Equilibrium Model. *Transportation Research*, 31B, 209-223.

Richardson, A.J. and Taylor, M.A.P. 1978. A Study of Travel Time Variability on Commuter Journeys. *High Speed Ground Transportation Journal*, 12, 77-99.

Shankar, V. and F. Mannering. 1998. Modeling the Endogeneity of Lane-mean Speeds and Lane-speed Deviations: A Structural Equations Approach. *Transportation Research*, 32A, 311-322.

Sheffi, Y. and W.B. Powell. 1982. An algorithm for the equilibrium assignment problem with random link times. *Networks*, 12(2), 191-207.

Smith, M.J. 1979. The Existence, Uniqueness and Stability of Traffic Equilibria. *Transportation Research*, 13B, 295-304.

Taylor, M.A.P. (1982) Travel Time Variability - the Case of Two Public Modes. *Transportation Science*, 16, 517-521.

Uchida, T. and Y. Iida. 1993. Risk Assignment: A New Traffic Assignment Model Considering the Risk of Travel Time Variation. *Proceedings of 12th Transportation and Traffic Theory*, Elsevier, Amsterdam, 89-105.

Wardrop, J.G. 1952. Some Theoretical Aspects of Road Traffic Research. *Proceedings of the Institution of Civil Engineers*, II(1), 325-378.

Zhang, N. and W.H.K. Lam. 2001. An Alternative to the Road Network Reliability by Travel Demand Satisfaction Ratio. *Proceedings of the 4th International Conference on Management*, Xian Jiaotong University, China, May 5-7, 47-53.

Zhang, N. and W.H.K. Lam. 2002. The Reliable User Equilibrium Problem in A Stochastic Transport network. *Proceedings of the 7th Conference of the Hong Kong Society for Transportation Studies*, December 14, Hong Kong, 119-126.

9 Reliability Analysis of Road Networks and Preplanning of Emergency Rescue Paths

Yanyan Chen[1], Michael G.H. Bell[2] and Ioannis Kaparias[3]

[1] Transportation Research Center, Beijing University of Technology, China, Email: cdyan@bjut.edu.cn
[2] Centre for Transport Studies, Imperial College London, UK, Email: mghbell@imperial.ac.uk
[3] Centre for Transport Studies, Imperial College London, UK, Email: ik00@imperial.ac.uk

9.1 Introduction

Emergency vehicles require good access to residences after a large scale disaster. In the Great Hanshin-Awaji Earthquake of January 17, 1995, in the Kobe area of Japan, the access of emergency vehicles to the sites where they were needed was severely hindered by the traffic chaos in the aftermath of the earthquake, obstructing lifesaving activities. The urgent need for improving the reliability of road networks, and especially the reliability of emergency pathways, was thus confirmed.

When a great disaster, such as an earthquake, occurs, finding a fast and reliable path for the rescue vehicles to each emergency site is vital, in order to reduce indirect losses. Also, a reliable path away from the emergency site is needed in this case for emergency evacuation. As traffic data (e.g. road travel times) may deviate greatly from their normal values at the time of the disaster, the task of finding an optimal rescue route rapidly and precisely becomes very complicated for the transport manager or rescue organisation responsible. The difficulty not only lies in the lack of dynamic traffic data, due to possible damage of the real-time traffic information collection and transmission hardware, but also in the fact that the computation of routes over large areas is time-consuming, especially when many requirements have to be met at the same time. Pre-planning of optimal rescue routes is an efficient way of reducing the emergency response time, as well as the reliance on real-time post-disaster data.

The aim of this chapter is to present an efficient method for pre-planning emergency paths, so as to reduce the rescue response time and to provide guidelines for disaster preparations. Using static link travel time estimation and network reliability analysis, an approach to emergency pre-planning based on a partially disjoint candidate path set for each important O-D (origin node-destination node) pair is proposed. A heuristic link weight increment method for finding partially disjoint path sets is introduced. By imposing constraints, it is ensured that every candidate path is both fast and reliable. Also, the probability of joint failure of all candidate paths is reduced by introducing constraints on the extent of path overlaps. Finally, the efficiency of the pre-trip path computation is improved by choosing a better heuristic function in the A* algorithm.

The outline of this chapter is as follows. In Section 2, previous work on this topic is reviewed. Section 3 describes the method of network connectivity assessment under failure independence, while Section 4 presents the methodology developed in order to perform reliability analysis taking failure dependence into account. Section 5 describes the heuristic method developed in order to obtain a partially disjoint set of candidate emergency paths. The new algorithm is tested on randomly generated road networks and the numerical results are presented in Section 6. Section 7 discusses the conclusions.

9.2 Background

9.2.1 Transportation Network Connectivity Reliability

In evaluating transportation network reliability, the flow (of people, vehicle, goods, etc.) may have a normal or abnormal state. Transportation network reliability has two forms; *connectivity reliability* and *travel time reliability*. Connectivity reliability is the probability that a given destination can be reached at all, while travel time reliability is the probability that a given destination can be reached within an acceptable amount of time. Many studies estimate transportation network reliability for the major road network within a large area, e.g. at the level of intercity highways (Iida and Wakabayashi, 1990; Bell and Iida, 1997). As the connectivity of the road network is the basic requirement for emergency vehicles to have access to residences and other important areas after an earthquake, connectivity reliability can be used to estimate the adequacy of the candidate emergency path set. Taking into account that non-emergency vehicles can be prohib-

ited from using emergency paths under emergency conditions, the flows on the paths are ignored when computing connectivity reliability.

The methods available for network connectivity reliability analysis can be classified into three categories; exact methods, heuristic methods and simulation methods. Exact methods include the enumeration method, the inclusion-exclusion formula, Fratta and Montanari's method, the path-and-cut method and others (Bell and Iida, 1997). The exact calculation, however, requires Boolean algebra and may involve extensive calculation, since both computation time and memory requirements increase exponentially with the number of links. To deal with such problems, heuristic methods that find the best upper and lower bound of reliability are developed. A simple method involves not making use of all paths and omitting Boolean algebra. When the number of paths is limited, an approximate reliability value can be obtained with a small amount of processing (Bell and Iida, 1997). Nevertheless, for large networks, such as transportation networks in urban and rural areas, a frequently used practical method is Monte Carlo simulation (Kumamoto, 1977). Even with this method, many special techniques are needed to obtain accurate values, while the amount of calculation is large.

Regarding transportation network reliability in abnormal situations, such as earthquakes, a number of studies consider the emergency evacuation case (Cova and Church, 1997; Church and Cova, 2000; Cova and Johnson, 2003), while others attempt to identify critical paths in the network, i.e. paths which minimise the total travel time while at the same time maximising the population served (Viswanath and Peeta, 2003). The economical impact of an earthquake on transport infrastructure, expressed as the cost of retrofitting bridges located on critical links, is also considered (Viswanath and Peeta, 2003; Sohn et al, 2003).

Nevertheless, many issues still remain to be solved. For example, mutual relationships or dependences among links are generally overlooked in studies of the seismic reliability of transportation networks. Some links may fail together in the same earthquake and are therefore strongly failure dependent. Thus, in network reliability analysis it is desirable to take the correlation between link failures into account. Moreover, as transportation networks can be large and can have complicated topologies, seismic reliability analysis may require a large amount of computation. Therefore practical analytical methods for transportation network seismic reliability analysis need to be developed.

9.2.2 Shortest Path Algorithms

Travel time is the main concern in emergency transport, hence paths that detour too much are of little practical use. Consequently, short and fast paths are mostly preferred by the emergency services after an earthquake. In this chapter we refer to cost, which could be a weighted sum of time and distance, but is more likely to be just time. When we refer to a shortest path we in fact mean a minimum cost path. The shortest path search is a necessary sub-procedure in the calculation of the candidate emergency paths. In static networks, an efficient algorithm for the one-to-all shortest path problem (from one node to every other node in the network) is Dijkstra's algorithm (1959). However, for the one-to-one shortest path problem (from one node to another node), the leading algorithm is A* (Hart et al, 1968).

The A* algorithm uses an estimate of the minimum distance from each node to the destination to determine how likely it is that any node lies on the best route. The cost associated with a node n is $f(n) = g(n) + h(n)$, where $g(n)$ is the cost of the best path found so far from the start to node n, and $h(n)$ is an estimate of the minimum cost to the destination from node n. At each point, the node with the lowest f value is chosen for expansion. Provided $h(n)$ is an under-estimate of the minimum cost from n to the destination (like the cost based on the Euclidean distance), the A* algorithm will find the optimal solution. When $h(n)$ is replaced by 0, the algorithm reduces to Dijkstra's algorithm. The efficiency of the search increases as $h(n)$ approaches the actual value. Chabini and Lan (2002) further extend the A* algorithm to shortest path problems in time-dependent networks.

9.2.3 Multiple Shortest Paths Algorithms

Taking into account the randomness of earthquakes and their damage to the transportation system, more than one candidate emergency path has to be pre-planned. Traditionally in graph theory, two categories of algorithms exist in order to compute alternative paths; the k-shortest paths algorithms proposed by Eppstein (1998), Martins et al (1999) and Jimenez and Marzal (1999), and the totally disjoint paths algorithms proposed by Dinic (1970) and Torrieri (1992). Furthermore, in the related research topic of designing survivable networks, algorithms ensuring that a certain number of disjoint paths exist between a pair of nodes have been developed (Monma and Shalcross, 1989; Kerivin and Mahjoub, 2005).

These algorithms have some drawbacks when used for path pre-planning. For the *k*-shortest paths algorithms, after *k* paths have been constructed, a path-checking procedure has to be performed to choose those alternative paths that are acceptable, i.e. that satisfy certain acceptability constraints. Consequently, when the number of paths from a source to a destination is large, *k*-shortest path algorithms become very inefficient for selecting alternative paths. Moreover, alternative paths fail simultaneously when shared links fail. For the disjoint path algorithms, on the other hand, the final path set is totally disjoint and the probability of all candidate paths failing simultaneously is significantly reduced. This nevertheless occurs at the cost of optimality, as the primary shortest path of the network may not be included in the set or the duration of many paths may be so long, that time constraints imposed by the emergency conditions are not met.

Relatively recently, several studies have been carried out in order to determine *k* partially disjoint paths in a network, subject to a number of constraints. The algorithm developed in this chapter also falls into this category. A method for finding *k* partially disjoint paths is the link weight increment algorithm, the idea of which arises from the following fact: in a shortest path algorithm, like Dijkstra or A*, a link *i* with a greater weight w_i will have a lower probability of being included in the shortest path P_{st} from node *s* to node *t*. If we set $w_i = \infty$, then link *i* will never be included in P_{st}, provided alternative paths with less than infinite length (or duration) are available. This can also be achieved by deleting link *i* from the network. Rouphail et al (1995) developed a pre-trip path planning scheme for providing tourists with alternate routes by increasing each link weight in P_0 (the shortest path) by 20%, 50% and 100% of its original value. Pu et al (2001) extended this approach to seek alternate paths that have a minimum number of double- and triple-shared links with the other paths obtained thus far, using Dijkstra's shortest path algorithm and a logarithmic link weight increment procedure.

Link weight increment algorithms can make the candidate emergency paths share as few links as possible, which in turn reduces the probability of joint failure of the candidate paths. The term "path failure" indicates that the path becomes unusable while the term "joint failure" of the candidate paths means that the candidate paths fail together. Without doubt, joint failure of the candidate paths should be avoided as much as possible. Unfortunately, the existing literature does not present any attempt to either minimise the probability of joint failure of the candidate paths or to avoid lengthy detours.

9.3 Reliability Assessment with the Disjoint Sub-path Method

9.3.1 Network Connectivity Reliability Analysis Under Failure Independence

A path is a series of links from the origin node to the destination node. When any link on a path is deleted, it is no longer connected. In transportation networks under emergency conditions, it can be assumed that the most direct paths available will be chosen as emergency paths, which will therefore by acyclic. Therefore, the term "path" in this chapter refers to acyclic paths.

The connective reliability of an O-D (origin-destination) pair is defined as the probability that there exists at least one connected path for this O-D pair during a specified period. Supposing that the conditions of failure for different links as well as routes are statistically independent from each other, the reliability of the *n*th O-D pair can be estimated using the following formula:

$$\phi_S^n = P\{\bigcup_{i=1}^{m} A_i\} \tag{9.1}$$

where ϕ_s^n is the seismic connectivity reliability of the candidate path set for the *n*th O-D pair and A_i represents the connectivity event of the *i*th path in the candidate path set, which has *m* elements.

By converting the logical summation into an algebraic summation (Fratta and Montanari, 1973), the union term can be expressed as the sum of disjoint path events:

$$\bigcup_{i=1}^{m} A_i = A_1 + \overline{A_1} A_2 + \cdots + \prod_{i=1}^{m-1} \overline{A}_i A_m \tag{9.2}$$

where the product operator "$\prod$" represents the joint operation "$\cap$", and the summation operator "+" represents the algebraic summation.

If the *m* paths do not overlap, this equation can be applied directly, thereby reducing the amount of calculation in relation to the enumeration method. If overlapping occurs, Boolean algebra is required, making the calculations more complicated.

Supposing that B_i, B_j and B_k are Boolean functions Liao (1982) proved the following disjoint sum principles:

$$\overline{B}_j B_k = \overline{B}_{j \leftarrow k} B_k \tag{9.3}$$

$$\overline{B}_i \overline{B}_j B_k = \begin{cases} \overline{B}_{i \leftarrow k} B_k & if \quad B_{j \leftarrow k} \subset B_{i \leftarrow k} \\ \overline{B}_{j \leftarrow k} B_k & if \quad B_{i \leftarrow k} \subset B_{j \leftarrow k} \\ \overline{B}_{i \leftarrow k} \overline{B}_{j \leftarrow k} B_k & \text{otherwise} \end{cases} \tag{9.4}$$

where $B_{j \leftarrow k}$ indicates the joint event of the units in *j* but not in *k* functioning normally (the bar indicates not functioning rather than functioning).

According to the above disjoint sum principles and De Morgan's Law in set theory, the union term can be further represented as

$$\bigcup_{i=1}^{m} A_i = A_1 + \sum_{j=2}^{m} \left(\prod_{1 \le i \le j-1} \overline{A}_{i \leftarrow j} \right) A_j = \sum_{i=1}^{M} A_{dis,i} \tag{9.5}$$

where $A_{i \leftarrow j}$ is the joint state of the units left on path *i* after deleting the units on path *j*, $\overline{A_{i \leftarrow j}}$ is the complementary state of $A_{i \leftarrow j}$, $A_{dis,i}$ is the joint state of the units on disjoint sub-path *i*, $A_{dis,1}$ is A_1 and M is the number of disjoint sub-paths in the sub-path set. Note that $M \ge m$. Then

$$\phi_S^k = P\left\{ \bigcup_{i=1}^{m} A_i \right\} = \sum_{i=1}^{M} P\{A_{dis,i}\} = \sum_{i=1}^{M} \left[\prod_{j_{1i}=1}^{J_{1i}} r_{j_{1i}} \prod_{j_{2i}=1}^{J_{2i}} \left(1 - r_{j_{2i}}\right) \right] \tag{9.6}$$

where j_{1i} is the *j*th unit in a normal state on the *i*th disjoint sub-path, $r_{j_{1i}}$ is the reliability of the *j*th unit on the *i*th disjoint sub-path, J_{1i} is the number of units in a normal state on the *i*th disjoint sub-path , j_{2i} is the *j*th failed unit on the *i*th disjoint sub-path, $r_{j_{2i}}$ is the reliability of the *j*th unit in a failed state on the *i*th disjoint sub-path and finally J_{2i} is the number of units in a failed state on the *i*th disjoint sub-path.

For the example network *Q* shown in Fig. 9.1 (only link reliability is considered), four acyclic paths from node 1 to node 4 are found:

$$A1 : [4] * [3]$$

$$A2 : [2]*[1]$$

$$A3 : [4]*[5]*[1]$$

$$A4 : [2]*[5]*[3]$$

where "*" represents the joint operation "$\cap$" and [i] represents the normal state of link *i.* Then,

$$A_{1\leftarrow 2} : [3]*[4],\ A_{1\leftarrow 3} : [3],\ A_{2\leftarrow 3} : [2],\ A_{1\leftarrow 4} : [4],\ A_{2\leftarrow 4} : [1],$$
$$A_{3\leftarrow 4} : [1]*[4] \subset A_{1\leftarrow 4}$$

As $A_{3\leftarrow 4} : [1]*[4] \subset A_{1\leftarrow 4}$, $A_{3\leftarrow 4}$ could be *absorbed*.

Using De Morgan's Law, expand and merge

$$A_1 + \sum_{j=2}^{m}\left(\prod_{1\le i\le j-1}\overline{A_{i\leftarrow j}}\right)A_j$$

to obtain the following disjoint sub-paths

$$A_{dis,1} : A1$$

$$A_{dis,2} : A2 * [-3]$$

$$A_{dis,3} : A2 * [3] * [-4]$$

$$A_{dis,4} : A3 * [-2] * [-3]$$

$$A_{dis,5} : A4 * [-1] * [-4]$$

where [-i] and [i] represent the failed and normal states of link *i* respectively.

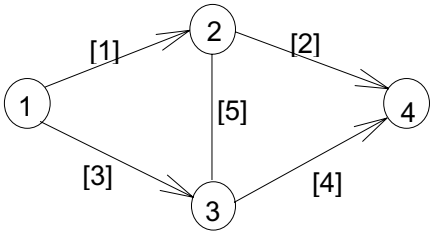


Fig. 9.1 Example network *Q*

This algorithm to calculate network reliability is called the disjoint path algorithm in this chapter.

9.3.2 Network Connectivity Reliability Analysis under Failure Dependence

The above algorithm assumes failure independence. But for road networks, especially when an earthquake occurs, failure dependence is an observed phenomenon. This means, when a link *i* fails other related links may also fail.

Determining the degree of failure dependence involves calculating the conditional probability of certain links failing, provided certain other links fail. Three types of failure dependence between links *i* and *j* are considered:

1. When link *i* fails but link *j* remains unaffected, then *i* and *j* are said to be *failure independent*.
2. When link *i* fails and link *j* fails with it, then *i* and *j* are said to be *totally failure dependent*.
3. When link *i* fails and the probability of link *j* failing with it lies between the probabilities of failure independence and total failure dependence, then *i* and *j* are said to be *partially failure dependent*.

The type and degree of failure dependence is represented by the coefficient μ_{ij}, which is the failure dependence coefficient of link *j* on link *i*. If *j* is failure independent from *i*, then $\mu_{ij} = 0$, whereas if *j* is totally failure dependent on *i*, then $\mu_{ij} = 1$. If *j* is partially failure dependent on *i*, then $0 < \mu_{ij} < 1$, with values of μ_{ij} increasing with stronger degrees of failure dependence.

It has been proven that the system reliability value, under the condition of partial failure dependence, lies between the system reliability value under the condition of failure independence (lower bound) and the system re-

liability value under the condition of total failure dependence (upper bound). However, the interval between the lower bound and upper bound increases rapidly with increasing number of units. Therefore, several research studies have concentrated on determining better bounds.

In this chapter, we use an extreme method, in which the units with failure dependence coefficient beyond a given threshold are classified to a failure dependence group, and each unit in the group is assumed to be totally failure dependent. And then each failure dependence group is taken as a new compound unit and is numbered accordingly. When calculating the network reliability by formula (6), the numbers of all units of each failure dependent group are replaced with the number of the corresponding compound unit. By this method, network reliability could be calculated approximately.

Suppose there are N important O-D pairs in the system, then the system reliability can be estimated using the following equation:

$$\phi_S = \sum_{k=1}^{\mathrm{N}} \xi_k \psi_{\mathrm{S}}^{\mathrm{k}} \tag{9.7}$$

In this equation, ξ_k is a parameter representing the importance of the kth O-D pair with $\sum_{k=1}^{\mathrm{N}} \xi_k = 1$. For a large network, the most efficient exact method is to find all the acyclic paths and disjoint them. As the number of candidate emergency paths for each O-D pair is limited, such a disjoint path algorithm is practical. Therefore, in this chapter the disjoint path algorithm considering failure dependence mentioned above is used to calculate the reliability of the sub-network consisting of candidate emergency paths.

9.4 Road Network Seismic Reliability Assessment

9.4.1 Unit Seismic Reliability Analysis

An urban transportation network can be regarded as a system comprising a set of links (road segments) and a set of nodes (intersections). Each component (link or node) is called a *unit*. By a binary assumption, units are presumed to have two states; functioning or failed. In its functional state, the unit is accessible by the traffic, whereas in its failed state, traffic cannot access the unit because of rubble from collapsed houses on the street or

because of damage. The reliability of the unit is the probability that it is functional during a specific period. In this chapter, the reliabilities of both links and nodes in urban networks are considered for the seismic reliability analysis.

To determine the reliability of the urban transport system, the reliability of the units involved must first be determined. In urban areas, links fail because rubble from neighbouring collapsed buildings falls on the streets and blocks roads. According to Li and Tsukaguchi (1997), the extent of road blockage in a street segment of an urban area after an earthquake of a certain magnitude is mainly affected by two factors; the street width and the proportion of the length of the street with wooden housing along it. Link reliability can be obtained through analysis of historical data. Lee and Yeh (2000) developed a discrete model with dummy variables in which the explanatory variables are the street width, the number of floors of the neighbouring buildings, the structure of the buildings, the street wall length, the utility poles and the street lights. Other researchers estimated link reliability through the possible "collapse area" of the neighbouring buildings. Any feasible method can be chosen to evaluate link reliability, according to data availability.

In an urban area, bridges and crossroads can be taken as units; the reliability of a bridge can be estimated either from historical statistical data, according to failure influence factors, or by dynamic analysis of the bridge structure (Zhao, 1994). In this chapter, the node reliability for each turning movement is assumed to be the same. For easily restored intersection nodes and road segments, a reliability value of 1 could be taken for simplification.

9.4.2 Unit Seismic Failure Dependence Analysis

During an earthquake, some links may fail together from the same cause. This is a special case of failure dependence, which is defined as "common cause failure".

As transportation networks cover large areas, structures with different seismic loading capacities founded in different types of ground have to be dealt with. Therefore, failure dependence coefficients between links are difficult to estimate in an analytical way. It is thus necessary to find an approximate method to estimate the degree of failure dependence between links.

As discussed above, the failure of units in an urban transportation network when an earthquake occurs is mainly caused by the collapse of

neighbouring buildings and by the collapse of bridges. Usually there are two causes of damage to buildings and bridges:

1. *Environmental failure* occurs from faulting or liquefaction in the ground on which the roads and buildings are founded.
2. *Structural failure* implies that damage to the structures occurs because the impact of the earthquake is beyond their designed capacity.

The failure dependence caused by environmental failure is thus called *environmental failure dependence* and that caused by structural failure is called *structural failure dependence*.

From historical data, whenever in an earthquake a fault breaks or liquefaction occurs in an area, almost all units in this area will fail together. Also, the probability of environmental failure is higher than the probability of structural failure of the neighbouring buildings and bridges. Therefore, it can be assumed that the units affected by the same fault or liquefied area are totally failure dependent.

The structural failure dependence condition is more complicated. Some units have a high probability of failing simultaneously in an earthquake and therefore exhibit strong failure dependence. Other units may fail simultaneously in one earthquake but not in another, therefore exhibiting weak failure dependence. As determining exact values for failure dependence coefficients is difficult, the following principles are adopted, so as to approximately distinguish strong structural failure dependence from weak structural failure dependence.

The seismic performance of bridges depends on their seismic design standard and their natural frequency of vibration. Bridges with same seismic design standard and natural frequency of vibration could be regarded as having a higher probability of failing together, and therefore would be strongly failure dependent in the same earthquake. On the contrary, bridges with different seismic design standards or different natural frequencies of vibration are weakly failure dependence in the same earthquake.

According to the above analysis, the following steps are taken in seismic reliability analysis for failure dependence in a road network:

1. The task of determining failure dependence is divided into two subtasks, namely determining environmental failure dependence and structural failure dependence.
2. It is assumed that the environmentally failure dependent units and the strongly structurally failure dependent units are totally failure dependent; it is also assumed that the weakly structurally failure dependent units are failure independent.

3. The links are classified, regarding the total failure dependence assumption, into different failure dependence groups according to their failure dependence causes.

More specifically, the following principles for the grouping of units into categories of failure dependence apply:

1. All units with a reliability value lower than a threshold are placed into a failure dependence group, the reliability of which is the average of the reliability values of all the units in the group. For all remaining units in the network, grouping follows the next principles.
2. The units that lie on the same fault or on the same area of sand prone to liquefaction are placed into another failure dependence group. The failure probability of this group is equal to the break probability of the fault or the probability of liquefaction of the sand.
3. The units with the same seismic design standard and similar natural frequency of vibration are placed into another failure dependence group. The reliability of the group is that of the unit with the minimum reliability value in the group.

Each group is considered as a new compound unit, with all the compound units and all the units not belonging to any of the above groups being taken as failure independent units.

9.4.3 Seismic Network Connectivity Reliability Analysis Under Failure Dependence

Having determined failure dependence relationships, the seismic connectivity reliability of the transport system under failure dependence can be found next, using the disjoint path algorithm as follows:

1. A failure dependence analysis for all links in the network is carried out and failure dependence groups are formed.
2. The least weight (fastest) path between the origin and destination nodes is found.
3. To take failure probability and dependence into account,
   - Each failure dependence group is taken as a new compound unit and is numbered accordingly. The numbers of all units of each failure dependent group are replaced with the number of the corresponding compound unit.
   - Generate a partially disjoint candidate path set which seeks to avoid unreliable units and compound units.
4. Constraints are applied on the candidate paths.

5. Seismic connectivity reliability is calculated by formula (5).
6. The network reliability value is computed assuming failure independence, using formula (6).

The key to the construction of an efficient candidate emergency path set is the threshold chosen for alternative paths. We select our candidate emergency paths according to two requirements; the *reasonable path requirement* and the *alternative path requirement*. The first requirement ensures that the selected paths are reliable and not too circuitous, while the second one ensures that joint failure on the candidate path set is unlikely to occur.

The reasonable path requirement is met by imposing the following constraints:

- **Constraint 1:** Maximum path duration
- **Constraint 2:** Minimum path reliability

The reliability of a path can be defined as its probability of being connected during a given period. This can be estimated by considering the path as a series of units.

In an earthquake, it is very likely that more than one unit will fail, implying that the candidate paths that contain failed units will fail together, causing joint failure. To improve the efficiency of the preparation, it is important to reduce the probability of joint failure of the candidate emergency paths. In this chapter, the reliability of the emergency candidate path set is used to quantify the probability of joint failure of the candidate path set. The reliability of the emergency candidate path set is defined as the probability of at least one candidate path being connected during an earthquake, which can be represented as follows:

$$\phi_S^n = P\left\{\bigcup_{i=1}^{K} A_i\right\} \tag{9.8}$$

where ϕ_s is the reliability of the candidate path set for O-D pair n and A_i represents the event that emergency vehicles can use the ith candidate path in the candidate path set, which has K elements. The reliability of the candidate path set can be computed using the disjoint path formula described above.

Generally, the higher the reliability of the candidate path set, the less the chance of all candidate paths failing together during an earthquake. Consequently, the alternative path requirement can be met by planning enough candidate paths. This is ensured by the following constraints:

- **Constraint 3:** Minimum reliability of candidate path set.

- **Constraint 4:** Minimum and maximum number of alternative paths.

A maximum number of alternative paths is set to avoid the use of too much computation time and storage space. For path sets with a low reliability value even when using the maximum number of emergency paths, a "strengthening" method to increase reliability should be developed.

Although the alternative path requirement is better met when all the candidate paths are totally disjoint, it is possible that a sufficient number of totally disjoint paths that meet the reasonable path constraints cannot be found. A candidate path set that is partially disjoint, meaning that the candidate paths are allowed to share some units, has to be sought in this case, limiting though the extent of sharing. Nonetheless, this relaxation makes the path set reliability computation more difficult than it is for simple parallel or series systems.

By satisfying the reasonable path constraints listed above, it can be ensured that the candidate paths are fast and reliable. By satisfying the alternative path constraints, the probability of all candidate emergency paths being blocked simultaneously is reduced.

9.5 A Heuristic Method for the Construction of the Candidate Emergency Path Set

9.5.1 Algorithm Description

Theoretically, the candidate path set should be constructed by enumeration. At first, a search through all the reasonable paths would be carried out and all possible candidate path sets with the minimum number of paths would be enumerated, before identifying the set with the highest reliability value. If the reliability value of this set did not meet the minimum reliability threshold imposed by the alternative path requirement, then the number of candidate paths would be increased and the enumeration and comparison steps would be repeated, until the alternative path requirement was met.

Nevertheless, such an enumeration algorithm would require long computation times because the number of possible candidate path sets increases exponentially with the number of reasonable paths. According to the concepts of joint failure and probability theory, an alternative path should avoid sharing or encountering high-risk units as much as possible; this can be achieved using link weight increment algorithms. In this sec-

tion therefore, a heuristic path searching and checking algorithm is suggested, in order to efficiently compute good candidate path sets. The algorithm is executed in two stages; in the first stage, reasonable paths are constructed using the weight increment procedure in conjunction with a shortest path algorithm, while in the second stage the paths obtained in the first stage are checked, with the aim of retaining only those that satisfy the imposed constraints. Every alternative path calculated is checked against both the reasonable path and the alternative path set constraints. The algorithm terminates when any of the alternative path constraints is satisfied. Using the weight increment procedure, the reliability of the path set and the reliability of candidate paths within the set are increased efficiently.

In this algorithm, the fastest path is taken as the default candidate path. At first, Dijkstra's algorithm is used to compute the fastest paths for all O-D pairs. Then, as different O-D pairs imply that the weights of different links should be increased, the alternative paths are computed for each O-D pair separately. The A* algorithm is used as a basic algorithm to compute an alternative path at each iteration.

The following is a description of the algorithm.

Notation

S_n	Candidate path set for O-D pair n;
w_i	Normal mean travel time of unit i;
r_i	Reliability of unit i;
W_0	Large value to be added to w_i;
w_i'	Incremented weight of unit i;
$P_{n,0}$	Shortest path for O-D pair n;
$L_{n,0}$	Normal mean travel time of $P_{n,0}$;
$P_{n,k}$	Candidate path k for O-D pair n; k=0,1,2,…;
$L_{n,k}$	Normal mean travel time of $P_{n,k}$;
$R_{n,k}$	Reliability of $P_{n,k}$;
$R'_{n,k}$	Lower reliability limit of $P_{n,k}$;
β	Upper limit of path normal travel time;
$L_{n,0}$	
$\phi_{n,S}$	Candidate path set reliability for O-D pair n;
$\phi_{n,S}$'	Lower reliability limit of candidate path set;
K_n	The number of candidate paths for O-D pair n;
N	Upper limit of candidate paths number.

Procedure

Step 0: Compute the duration of shortest paths for all O-D pairs

Set S_n = "empty" for all O-D pairs n. Calculate the least normal travel time paths and their duration for all O-D pairs using Dijkstra's algorithm. Save $P_{n,0}$ to S_n and set K_n =1.

Step 1: Iteration
For each O-D pair n, set iteration number m = 0 and execute Steps 2 to 5.

Step 2: Unit weight increment
Every unit on a path in S_n, every unit that is positively failure dependent on high-risk units on paths in S_n, and every other high-risk unit in the network have their weights increased by Δw_i,

$$w_i' = w_i + \Delta w_i = w_i + \alpha^m (1 - r_i)^q W_0 \tag{9.9}$$

where $0 < \alpha < 1$ and $q = 0$ when $m = 0$, otherwise $q = 1$. Increment K_n and m by 1.

Step 3: Find alternative candidate paths
Recalculate the least weight path $P_{n,k}$ based on w_i' using the A* algorithm. Then restore the unit weight to w_i and compute $L_{n,k}$, which is the sum of w_i in the path $P_{n,k}$.

Step 4: Path checking
If $P_{n,k}$ violates the constraint, i.e. $L_{n,k} < \beta L_{n,0}$ or $R_{n,k} < R'_{n,k}$, discard path $P_{n,k}$, decrement K_n by 1 and go to Step 2. Otherwise, save path $P_{n,k}$ to S_n and compute $\phi_{n,S}$.

Step 5: Termination
If $K_n < N$ or $\phi_{n,S} < \phi_{n,S}'$ go to step 2, else terminate.

To obtain an alternative path using this algorithm, the weights of all units that are included in paths in S_n and the weights of high-risk units are initially increased by W_0 (when m =0), which ensures that the alternative path computed avoids units on paths in S_n, units that are failure dependent on high-risk units on paths in S_n, and other high risk units in the network, provided such a path exists. However, if the duration constraint is violated, meaning that the path is too circuitous, then some units that were previously avoided, need to be re-included. The weight increment Δw_i is reduced and the least weight path is recalculated.

When $m \geq 1$, the function $\alpha^m (1 - r_i)^q$ ensures that Δw_i is reduced (because $0 < \alpha < 1$) with increasing iterations, but also that the weight of higher risk units is reduced less than the weight of lower risk units so as to ensure that higher risk units have a lower probability of appearing in the candidate path than do lower risk units.

In the algorithm, W_0 should be large enough to prevent any high-risk units from being selected in the first iteration. However, the efficiency of the algorithm would be improved if W_0 is as small as possible, by reducing the number of iterations. Generally, the more the alternative paths and the smaller the number of high-risk units in the road network, the smaller W_0 could be. It is suggested that W_0 should lie in the range 1.5 $L_{n,0}$ to 3 $L_{n,0}$.

In fact, to get a more exact solution, Δw_i should be "dither-changed", which means that when the constraint is met after a reduction of Δw_i, it should be increased again between the value of the last iteration and the present iteration, because there could be a path which better avoids the sharing of high-risk units, whilst still meeting the duration constraint. However, such a procedure would result in heavy computation, so in this chapter a monotone weight reduction strategy is used. Although it may not be optimal, it is faster and yields acceptable results.

In the implementation of this algorithm, α can be adjusted to balance the number of iterations and the accuracy of the solution. The bigger α is, the closer the solution gets to the optimum, but the greater the number of iterations and the slower the speed of execution of the algorithm.

For simplification, only the units that are strongly failure dependent on the high-risk units on the selected candidate paths are avoided.

As the reliability of an alternative path is supported by avoiding high-risk units, the reliability of a candidate path set is supported by meeting the minimum number of candidate paths. Of course the best minimum number of candidate paths could differ by area or O-D pair and can be determined by an offline analysis of the relationship between the number of candidate paths and the reliability of the path set.

9.5.2 Further Improvements to the Construction of the Candidate Path Set

As mentioned above, the heuristic function *h(n)* provides the A* algorithm with an estimate of the minimum cost from any vertex *n* to the destination. The choice of a good heuristic function is a critical step towards improving the efficiency of the algorithm. In transportation networks, to ensure that the solution obtained will be exact, *h(n)* is often taken as the Euclidean distance divided by the highest value of speed in the network. This guarantees that *h(n)* is less than (or equal to) the actual least cost from *n* to the destination.

However, when computing the alternative paths, the weights of some units on the shortest path between an O-D pair will have to be increased, so the new shortest path cost will be greater than (or equal to) the shortest

path cost before the unit weight increment procedure. Therefore, the shortest path cost for all O-D pairs before the unit weight increment procedure will be higher than the estimate based on the Euclidean distance, but lower than the actual shortest path after the unit weight increment procedure. Based on this fact in the present heuristic algorithm, after the shortest paths of all O-D pairs based on normal travel time data have been calculated by Dijkstra's algorithm, the travel times on the least travel time paths can be used as heuristics *h(n)* for the A* search in the subsequent alternative path computation for any O-D pair. The efficiency of A* will be improved as less nodes will be expanded while still ensuring that the optimal path will be found.

9.6 Numerical Experiment

The algorithm was tested for different randomly generated grid networks and conditions. First the results for a small network with 36 nodes and 60 links under three conditions are discussed, so as to illustrate the construction of a reasonable candidate path set. Then the results for a relatively large grid network with 2800 nodes are discussed so as to demonstrate the efficiency of the search for alternative paths when using the information obtained from the preceding shortest path computation.

The small network discussed in this chapter is shown in Figs. 9.2, 9.3 and 9.4. In the graphs, a node is represented by a circle and the node number is shown inside it. Links are represented by thin grey lines, while their lengths correspond to their airline distances. Normal mean speeds and reliability values are shown next to each link, where the normal mean speed is a value between 30 and 60 (km/h) and the reliability is a value between 0 and 1. The origin and destination nodes are represented by black rectangles. The upper limit for the number of candidate paths for each O-D pair is 3. Path 1 (the shortest path) is represented by a thick black line, path 2 is represented by a thick grey line and path 3 is represented by a thick dashed line.

In Fig. 9.2 no circuitous path constraint is considered and it can be seen that there are almost no shared links in the candidate path set. The travel time of path 1 is 36.0 min and its reliability is 0.67. Path 2 has a travel time of 39.2 min and a reliability of 0.56, while the corresponding values for path 3 are 50.3 min and 0.33 respectively. The system reliability is 0.9009.

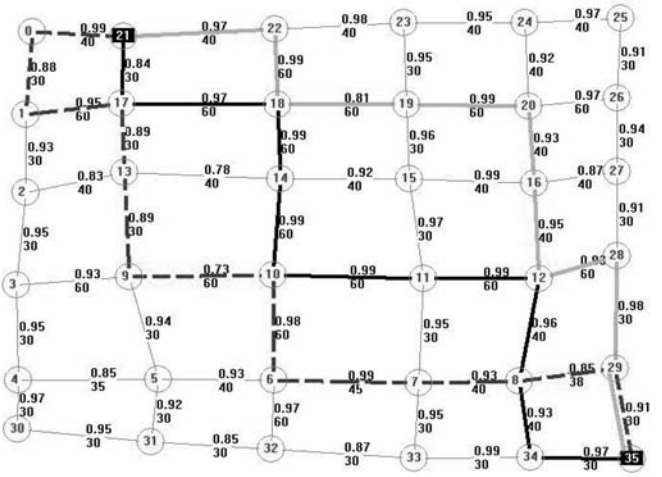


Fig. 9.2. Path set construction without circuitous path constraint

In Fig. 9.3 the circuitous path constraint is added with a circuitous permission parameter $\beta = 1.2$, but link reliability is not considered in the construction of the candidate path set. In this case, links will be shared as little as possible without violating the circuitous path constraint. The result shows that the travel time of path 1 is 36.0 min and its reliability is 0.67. Path 2 (same as path 2 in Fig. 9.2) has a travel time of 39.2 min and a reliability of 0.56, while the corresponding values for path 3 are 42.5 min and 0.33. The system reliability is 0.8844.

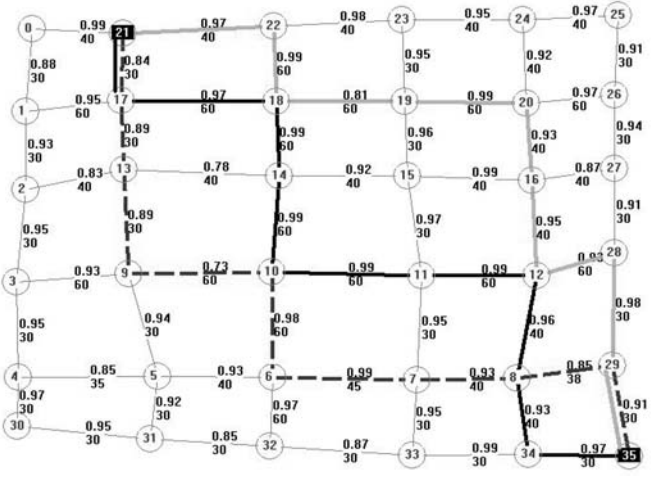


Fig. 9.3. Path set construction with circuitous path constraint (reliability not considered)

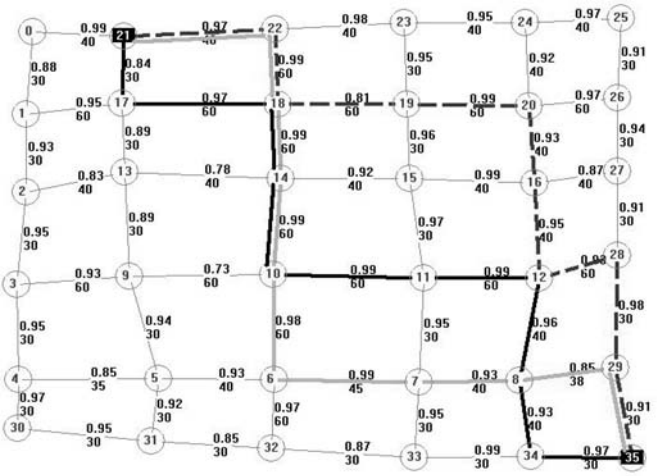


Fig. 9.4. Path set construction with circuitous path constraint, based on link reliability analysis

From the above results, it can be seen that although more links are shared in Fig. 9.4 than in Figs. 9.2 and 9.3, the system reliability is still higher than it is in either Fig. 9.2 or Fig. 9.3. The reason for this lies in the fact high-risk links are avoided in Fig. 9.4 even when the circuitous path constraint is considered, thereby improving system reliability.

To test the efficiency gain through the use of information obtained in the initial fastest path computation, the algorithm is tested on a randomly generated grid network with 2800 nodes. After the Dijkstra search for the shortest path for all O-D pairs, an O-D pair is randomly chosen and one of its alternative paths is computed. Figs. 9.5 and 9.6 show the results of the case where no previously computed information is used in the A* search and the case where it is, respectively. In the figures, the shortest path is represented by a black line and the alternative path is represented by a thinner line. The nodes explored are represented by thick black circles. The result shows that in case 1 (shown in Fig. 9.5), the number of explored nodes is 612, whereas in case 2 (shown in Fig. 9.6), the number of explored nodes drops to 81. Obviously the efficiency of the A* search for alternative paths is remarkably improved by using previously computed information.

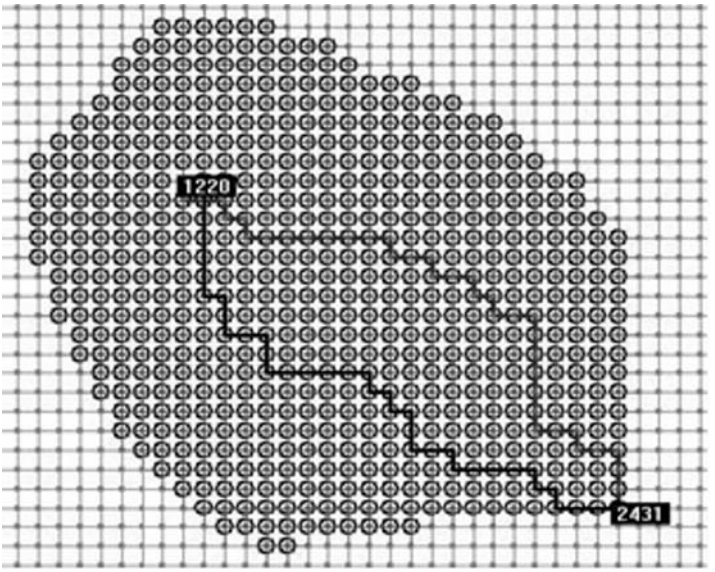


Fig. 9.5. No previous information is used

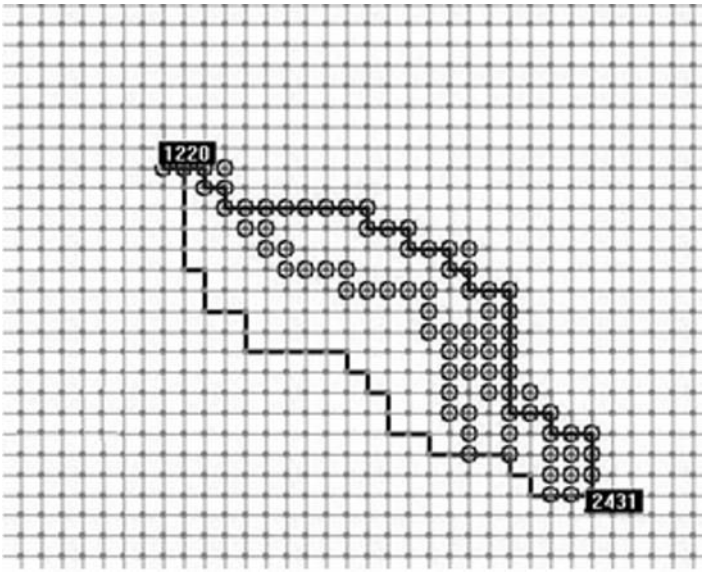


Fig. 9.6. Previous information is used

By numerous tests, the first alternative path can usually be obtained with up to 2 iterations, whereas the second alternative path can usually be obtained through 3 to 4 iterations. The denser the road network, the rarer

high-risk links and the more relaxed the circuitous path constraint, the less the number of iterations required.

9.7 Conclusions

Under emergency conditions in the aftermath of an earthquake, it is frequently the case that the shortest accessible routes cannot be found rapidly enough because of lack of exact traffic information for the entire network. This results in the post-earthquake rescue operations being delayed. Preplanning the emergency paths is an effective way to reduce rescue response time and to keep the post-disaster traffic under control. However, due to the occurrence of uncertain events, such as the blocking of roads and the collapse of bridges, it is difficult to derive reliable emergency routes in advance.

Based on seismic reliability and failure dependence analyses, a methodology is developed, according to which a partially disjoint candidate emergency path set can be constructed using a heuristic link weight increment method. The resulting paths satisfy both the imposed path length constraint, which ensures that circuitous paths are avoided, and the alternative path constraint, which limits the probability of joint failure of the candidate emergency paths. The availability of a candidate emergency path set provides the potential for quick, responsive rescue operations after an earthquake, as well as post-earthquake traffic management.

References

Ahmad SH (1982) A simple technique for computing network reliability. IEEE Transactions on Reliability, 31: 41-44

Bell MGH, Iida Y (1997) Network reliability. In: Transportation network analysis. Wiley and Sons, Chichester, pp. 179-192

Chabini I, Lan S (2002) Adaptations of the A* algorithm for the computation of fastest paths in deterministic discrete-time dynamic networks. IEEE Transactions on Intelligent Transportation Systems 3: 60-74

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numerische Mathematik 1: 269-271

Church RL, Cova TJ (2000) Mapping evacuation risk on transportation networks using a spatial optimization model. Transportation Research C 8: 321-336

Cova TJ, Church RL (1997) Modelling community evacuation vulnerability using GIS. International Journal of GIS 11: 763-784

Cova TJ, Johnson JP (2003) A network flow model for lane-based evacuation routing. Transportation Research A 37: 579-604

Dinic EA (1970) Algorithm for solution of a problem of maximum flow in a network with power estimation. Soviet Math. Dokl. 11: 248-264

Eppstein D (1997) Finding the k shortest paths. SIAM Journal of Computing 28: 652-673

Fratta L, Montanari UG (1973), A Boolean algebra method for computing the terminal reliability in a communication network. IEEE Trans. Circuit Theory 20: 203-211

Hart PE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics 4: 100-107

Huang Y-W, Jing N, Rundensteiner EA (1995) Route guidance support in intelligent transportation systems: an encoded path view approach. University of Michigan Technical Report

Iida Y, Wakabayashi H (1990) An approximation method of terminal reliability of road network using partial minimal path and cut sets. Proceedings of the 5th WCTR, vol. 4, pp. 367-380

Jimenez VM, Marzal A (1999) Computing the K shortest paths: a new algorithm and an experimental comparison. Lecture Notes in Computer Science Series 1688: 15-29

Kerivin H, Mahjoub AR (2005) Design of survivable networks: A survey. Networks 46: 1-21

Kumamoto H, Tanaka K, Inoue K (1977) Efficient evaluation of system reliability by Monte Carlo method. IEEE Transactions on Reliability 26: 311-315

Lee C-K (1994) A multiple-path routing strategy for vehicle route guidance systems. Transportation Research C 2: 185-195

Lee Y-L, Yeh K-Y (2000) A preliminary study of disaster prevention planning road function - Evaluation in Tainan City

Li GQ (1994) The anti-earthquake reliability analysis of civil building network system. Press of Earthquake (in Chinese), Beijing, China

Li Y, Tsukaguchi H (1997) Improving the reliability of street networks in high density populated urban areas. In: The reliability of networks, pp. 261-272

Liao JS (1982) The disjoint computation of network reliability. Journal of space navigation (in Chinese)3: 15-23

Martins EQV, Pascoal MMB, Dos Santos JLE (1999) Deviation algorithms for ranking shortest paths. International Journal of Foundations of Computer Science 10: 247-261

Monma CL, Shallcross DF (1989) Methods for designing communication networks with certain two-connected survivability constraints. Operations Research 37: 531-541

Pu J, Manning E, Shoja GC, Srinivasan A (2001) A new algorithm to compute alternate paths in reliable OSPF (ROSPF). Proceedings International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2001), pp. 299-304

10 Continuity in Critical Network Infrastructures: Accounting for Nodal Disruptions

Tony H. Grubesic[1], Alan T. Murray[2] and Jessica N. Mefford[2]

[1] Department of Geography, Indiana University, Bloomington IN 47405-7100, USA. Email: tgrubesi@indiana.edu
[2] Department of Geography, The Ohio State University, Columbus, OH 43210-1361, USA. Emails: murray.308@osu.edu; mefford.293@osu.edu

10.1 Introduction

Since the attacks of September 11 2001, there has been a renewed interest in identifying, protecting and maintaining the functionality of critical infrastructure systems in the United States and abroad. However, because these systems are relatively complex, many difficulties have emerged when attempting to differentiate between elements of the infrastructure that are "critical" and those that are not. In an effort to resolve this issue, the U.S. government has released several important reports detailing the specific infrastructures and assets considered critical to national security, governance, public health, and the economy (White House 2003). Table 10.1 displays the key sectors and assets outlined in the *National Strategy for the Physical Protection of Critical Infrastructures and Key Assets* (*Strategy*).

Table 10.1 Critical Infrastructure and Key Assets

Infrastructure	Assets
Agriculture and Food	National Monuments and Icons
Water	Nuclear Power Plants
Public Health	Dams
Emergency Services	Government Facilities
Defense Industrial Base	Commercial Key Assets
Telecommunications	
Energy	
Transportation	
Banking and Finance	
Chemicals and Hazardous Materials	
Postal and Shipping	

Source: (White House, 2003)

An important contribution of *Strategy* is the identification of five cross-sector security priorities that help frame a comprehensive plan for securing critical infrastructure. These priorities include: 1) planning/resource allocation, 2) information sharing and indicators and warnings, 3) personal surety, building human capital and awareness, 4) technology and research and development, and 5) modeling, simulation and analysis. Of particular interest in this chapter is the fifth cross-sector priority, modeling, simulation and analysis. *Strategy* notes that modeling and simulation can provide guidance in prioritizing the protection of critical infrastructure and key assets and is an important aspect to maintaining the continuity of their functions. However, *Strategy* also notes that:

> "the challenges and uncertainties presented by critical nodes and single-points-of-failure within infrastructures, as well as increasing interdependencies that exist among the various infrastructure sectors both nationally and internationally…. are often difficult to identify and resolve, as are the cascading and cross-sector effects associated with their disruption" (White House 2003, pp. 33).

More importantly, the report (*ibid*) acknowledges that modeling, simulation and analysis can:

"facilitate protection planning and decision support by enabling the mapping of complex interrelationships among the elements that make up the risk environment. For example, modeling traffic patterns through a particular junction, such as a key railhead or air terminal, allows analysis of the various possible outcomes of an attack on that node at various points in time" (pp. 33).

This clearly establishes the need to develop models that both mimic potential outcomes of an attack and provide a framework for prioritizing the protection of critical nodes in geographically linked networks (Cutter et al. 2003; Church et al. 2004; Grubesic and Murray 2006). In addition, it highlights the emerging importance of functional continuity in critical infrastructure during a potentially disruptive event.[1]

The purpose of this chapter is to examine the issue of continuity for geographically linked network infrastructures by evaluating the vulnerabilities of nodal interconnection points through the use of a spatial optimization model. The developed model provides an extension to the work of Grubesic and Murray (2006), in which the geographic impacts of losing vital nodes in geographically linked networks (e.g. telecommunication switching centers or electrical substations) were evaluated by measuring the overall disruption of flows along arcs. The new, bi-objective approach presented here offers additional insight into network continuity and the measurement of cascading failure by examining both node and arc attributes (e.g. population and capacity) simultaneously. As mentioned previously, the merits of this type of analysis are clear. First, modeling and simulation can provide guidance in prioritizing the protection of critical infrastructure and key assets and is an important aspect of maintaining the continuity of their functions. Second, a geographic approach to these types of problems provides an alternative viewpoint to the traditional, engineering-driven approaches focusing on component reliability and the probabilistic rates of failure attached to network arcs and nodes (Tillman et al. 1977; Mohamed et al. 1992; Kansal et al. 1995; Premkumar et al. 2000). Finally, as the importance of a geographic focus on critical infrastructure identification and protection continues to emerge, it is hoped that the approach introduced in this chapter, focusing on spatial relationships in a network, will provide a foundation from which future work in this area will be conducted.

[1] Continuity is broadly defined as the ability of a business, infrastructure provider or government agency to maintain operations and the provision of critical goods and services before, during and after a major disruptive event.

The remainder of this chapter is organized as follows. Section 10.2 provides additional background information on the spatial dimensions of critical infrastructure protection and cascading failure along with a brief review of the literature pertaining to network survivability and continuity. In the next section, we present an extension to the Node Removal Impact Problem (NRIP) and follow with some computational experience, presenting an example for a telecommunication network. Section 10.4 provides a summary of the results and is followed by a brief discussion and conclusions.

10.2 Background

The emergence of massive, interdependent, geographically linked, critical infrastructure systems in urban and suburban areas of the United States (and most of the developed world) coincides with the geographic diffusion of population to the urban periphery and the increasing density and vertical expansion of urban cores (O'Kelly and Horner 2003; Zimmerman, 2005). This is not to say that spatially complex and interlinked systems did not exist prior to these demographic and economic redistributions. However, the relative interconnectedness of these systems, particularly in urban centers, is increasing significantly (Coutard et al. 2005). In many respects, urban centers now represent a concentration of critical assets, such as telecommunication hubs, transportation hubs (air, rail, road), water distribution systems, electrical grids, banking and finance operations, medical facilities (Chang et al., 2002; Church et al. 2004; Church and Scaparra 2005; Grubesic and Murray 2006; Zimmerman 2005). As a result, the loss of vital hubs or distribution centers in any of the aforementioned infrastructures can have disastrous effects both within the existing system and across interlinked infrastructures. More specifically, the loss of a vital node can trigger the cascading failure of interconnected systems, producing a series of primary failures within an existing system or secondary failures in interdependent, linked infrastructures (Little 2002; Carreras et al. 2002; Albert et al. 2004; Talukadar et al. 2003; Houck et al. 2004; Grubesic and Murray 2006). This type of cascading failure was all too apparent during the North American electrical blackout of August, 2003. Although the power failure started in an electrical generation plant outside of Cleveland, Ohio, the regional impacts were widespread, impacting 50 million customers in eight U.S. states and two Canadian provinces. There were significant local impacts as well. For example, in New York City, 400,000 train commuters were stranded, 11,000 street intersections were without oper-

able signals, emergency 911 services were inoperable, water pumps failed and total economic losses were estimated at $1 billion dollars (Botelho 2003; Johnson and Lefebvre 2003). These localized impacts underscore the problems inherent to interconnected infrastructures. Not only does the loss of a vital node have the ability to impact the continuity of a geographically linked system, nodal loss in a critical infrastructure network can spawn additional problems throughout interdependent infrastructures.

Reliability or Survivability

Some of the most widely misunderstood aspects of critical infrastructure modeling are the subtleties between different measures of network performance. For example, while *reliability* and *availability* are often considered to be identical, these concepts are actually quite different. The same can be said for survivability, vulnerability, performability and continuity. Let us start with the basics, using telecommunication networks as an example. The U.S. Department of Commerce (1996) provides one of the more widely accepted definitions of telecommunication survivability:

> ***Survivability***: A property of a system, subsystem, equipment process, or procedure that provides a defined degree of assurance that the named entity will continue to function during and after natural or man-made disturbance; e.g., nuclear burst. Note: for a given application, survivability must be qualified by specifying a range of conditions over which the entity will survive, the minimum acceptable level or post-disturbance functionality, and the maximum acceptable outage duration.

As noted by Grubesic and Murray (2005), this definition is probabilistic in nature, requiring some level of assurance (i.e. probability) that a system will continue to function under duress. On the other hand, network reliability is a more broad-based term concerning the capability of the network to provide the connections required for network functionality (Colbourn 1999). More specifically, the most common measure of network reliability is connectedness, the probability that all terminal nodes in a network are reachable from all source nodes in the network. Similarly, network performability measures deal with the probabilities attached to both statistically dependent (e.g. environmental or overload) and independent (e.g. wearout) failures. Availability is a more subtle measure of both performability and reliability, which is the ability of a device (e.g. router, network link, etc.) to provide network services if and when they are needed, with-

out delay. Finally, the all-encompassing concept of continuity refers to the ability of a network to provide alternative modes of operation for those activities or processes which, if they were to be interrupted, might otherwise bring about a seriously damaging or potentially fatal loss to the network.

Previous research in the operational continuity of systems has both explicitly and implicitly addressed the overall functionality of networks under duress (Wollmer 1964; Corley and Chang 1974; Ratliffe et al. 1975; Colbourn, 1999; Medhi, 1999; Jrad et al. 2004; Boedhihartono and Maral 2003). While the majority of this research focuses on the probabilistic nature of network element failures, there is an overriding concern with the topological characteristics of a system. For example, in a given telecommunication network, what is the most "vital link"? Phrased somewhat differently, which link in the system, if lost, will create the most significant disruption to traffic or overall functionality? More importantly, what populations or businesses will be impacted if a particular link or node is lost? Unfortunately this type of scenario is not uncommon. For example, Pakistan recently suffered the loss of its only fiber-optic link to the Internet, an undersea cable that failed in June, 2005 (Reuters 2005). In addition to 10 million Internet users hit by the initial crash, only 20% of the normal capacity for international phone calls was available (Faisal 2005). Fortunately, about 50% of the Internet links to the country were restored using a satellite backup system. However, many of the more high-capacity, or "critical", links needed by the burgeoning call center industry in Pakistan remained without service for quite some time (Faisal 2005). In fact, a call center operator in the city of Karachi lost two of their most important U.S. clients due to the fiber failure (ibid). In this case, a lack of fiber diversity and redundancy had a detrimental impact on Pakistan. Conversely, if a fiber link connecting India, a major participant in the call center industry, had been damaged, there would still be four fiber paths available for rerouting critical transmissions (Faisal 2005). This example illustrates the problems and prospects associated with continuity in geographically interlinked systems. As mentioned previously, continuity is not simply a question of component failure in a system, it is also concerned with the spatial characteristics of interconnection and the ability to maintain operation when portions of the system fail. In addition, the ability to assess the population impacted by a failure is also critically important. The next section introduces an approach for measuring and modeling potential impacts of losing critical infrastructure elements, nodes and arcs that are geographically linked.

10.3 Modeling Potential Impacts

One important aspect of modeling potential damage to critical infrastructure is the ability to account for arc attributes. As mentioned previously, arcs are the connective links between nodes, serving as conduits for the flow of critical/valuable resources. For example, arcs (e.g. fiber optic backbones) carry information on telecommunication networks. Arcs also carry natural gas on pipeline networks and food and supplies via road and rail networks. Therefore, the ability to gauge potential damage to these conduits, particularly in terms of their location and diminished capacity or operation, is a major concern. Similarly, the ability to account for node attributes is important for modeling potential damage to critical infrastructure. Nodes often serve as the distribution points on a network, where flow is aggregated and redirected to an end destination. For example, airline hub-and-spoke networks rely on nodes (i.e. hubs) to aggregate and redistribute passengers in a cost efficient and timely manner. Backbone routers on the Internet serve a similar purpose, where bit packets are aggregated and redirected for information flows. However, in the context of geographically linked systems, nodes can also be conceptualized as large pools of end-users. In this context, modeling potential damage could include the population of an entire city (i.e. node) or a subset of the population. For example, the cascading failure of the Northeastern Interconnection of the North American Power grid impacted the entire population of Cleveland and Toronto, while it impacted a smaller subset of the population in New York and Columbus (Grubesic and Murray 2006).

From a methodological perspective, one approach for modeling the potential impacts of a major nodal disruption is the Node Removal Impact Problem (NRIP) introduced by Grubesic and Murray (2005). However, this approach only accounts for arc attributes, measuring the overall network damage associated with a nodal failure. The following approach focuses on both arc and node impacts.

Consider the following notation:

i =index of nodes
j =index of arcs
Ω_j =node pair defining arc j
a_j =attribute of arc j
c_i =attribute of node i

N_i =set of arcs incident to node i

n_i = number of arcs incident to node $i = |N_i|$

p =number of nodes to remove from network

$$x_i = \begin{cases} 1 & \text{if node } j \text{ is removed from network} \\ 0 & \text{otherwise.} \end{cases}$$

$$y_j = \begin{cases} 1 & \text{if arc } j \text{ is in network} \\ 0 & \text{otherwise} \end{cases}$$

As noted previously, node and arc attributes can take a variety of forms, depending on the nature of the network. For arcs, demand characteristics such as bandwidth, capacity or traffic volume are represented using a_j. For nodes, demand characteristics such as population or passengers are represented by c_i.

A model for addressing impacts associated with both arc and node failure/disruption is the following:

Bi-Objective - Node Removal Impact Problem with Node Attributes ***(BO-NRIP)***

$$\text{Maximize or Minimize } Z^1 = \sum_j a_j y_j \tag{1}$$

$$\text{Maximize or Minimize } Z_2 = \sum_i c_i x_i \tag{2}$$

Subject to

$$\sum_i x_i = p \tag{3}$$

$$\sum_{j \in N_i} y_j \le n_i (1 - x_i) \qquad \forall i \tag{4}$$

$$y_j \ge 1 - x_i - x_l \qquad \forall j \;\&\; (i,l) \in \Omega_j \tag{5}$$

$$y_j = \{0,1\} \qquad \forall j \tag{6}$$

$$x_i = \{0,1\} \qquad \forall i$$

There are two objectives in BO-NRIP, and hence the bi-objective designation. Objective (1) is to either minimize or maximize the total arc flow

impacted by the removal of nodes. Objective (2) is to either minimize or maximize the total nodal demand impacted by removed nodes. Constraint (3) specifies that p nodes are to be removed from the network. Constraints (4) stipulates that arcs cannot be included unless the associated end nodes are maintained in the network. Alternatively, Constraints (5) force incident arcs to be included in the network if end nodes are not removed. Collectively, Constraints (4) and (5) ensure that both maximization and minimization can be applied to the objective. Finally, integer requirements are imposed in constraints (6).

There are a number of interesting characteristics of BO-NRIP worth further discussion. The first objective of BO-NRIP, (1), is equivalent to that of NRIP, as are Constraints (3)-(6) (Grubesic and Murray 2006). Thus, NRIP is a special case of BO-NRIP. The uniqueness of BO-NRIP is the addition of objective (2). This objective is a measure of demand impacted within the networked system, reflecting nodal service disruption. However, instead of demand/flow along an arc, (2) is constructed to represent demand at node locations. In this particular instance c_i represents the population of node i (or city) in the network. As noted previously, in addition to helping estimate potential damage during a major disaster, node attributes are also an important factor in optimizing disaster mitigation efforts. For example, how might one prioritize mitigation efforts when nodal failures occur in a system? Phrased somewhat differently, what is the most efficient and cost-effective approach for restoring service in a system to attain maximum benefit? Although there are a variety of political, economic and environmental factors that influence decisions regarding systems defense or restoration, mitigation efforts would likely be devoted to areas where the greatest overall relief is possible to aid the greatest population. Finally, a unique aspect of this bi-objective formulation is that both objectives, (1) and (2), are subject to the same constraints, with additional interest focused on the characteristics of node i.

From an operational perspective, the loss of a node will render any arcs incident to that node inactive given the loss of connectivity in the BO-NRIP model. Therefore, if one is seeking to inflict *minimum* damage to the network, arcs will tend to be maintained if at all possible. However, if x_i=1, an end node of an arc is lost and Constraints (4) forces the deactivation of all arcs incident to that node. The reverse is true if one is seeking to inflict *maximum* damage to the network. Constraints (5) force the removal of network arcs. However, if x_i and x_l are equal to 0 (that is, they are not removed), then the arcs incident to these nodes remain active in the network.

There are four possible ways to approach BO-NRIP. Using the undirected, ICG fiber-optic backbone network as an example, each case can be illustrated. It is assumed, without loss of generality, that bandwidth reflects the link attribute of interest and population represents the node (or city) attribute of interest in the network. Thus, the four approaches for examining BO-NRIP are:

- min, min (minimize bandwidth available, minimize population impacted)
- min, max (minimize bandwidth available, maximize population impacted)
- max, min (maximize bandwidth available, minimize population impacted)
- max, max (maximize bandwidth available, maximize population impacted)

Thus, for BO-NRIP there are many combinations of optimization orientation.

It is also important to note that BO-NRIP is *simultaneously* addressing the two objectives. It is well known that bi, or multiple objectives represent a challenge in optimization, which is why much research has been devoted to such problems. A classic text on multi-objective optimization is that of Cohon (1978).

For planning, policy and decision making purposes, multi-objective models require a process for evaluating and establishing the relative importance for associated objectives being considered. Thus, methods have been developed for identifying non-dominated solutions and the associated non-inferior tradeoff curve. A popular approach reviewed in Cohon (1978) is the weighting method. The weighting method integrates the two objectives using a relative weight of importance, *w*. In the context of BO-NRIP we get:

$$wZ_1 + (1-w)Z_2 \tag{7}$$

By examining different values of *w* in a systematic way, one can identify non-dominated solutions. As noted previously, NRIP is actually a special case of BO-NRIP. For the weighting method one need only assign $w = 1$, giving all importance to arc attributes, thereby rendering BO-NRIP identical to NRIP. Alternatively, one could examine the case where *w* = 0, so all importance is now attached to node attributes. Other interesting tradeoffs are possible when $w \in (0,1)$. Thus, values of $w \in [0,1]$ are typically

evaluated in a systematic fashion in order to identify non-dominated solutions.

10.4 Case Study on the ICG Backbone Network

The empirical analysis was carried out on a Pentium 4/2.53 GHz personal computer. ArcView version 3.3 was employed to manage, manipulate and analyze a North American Internet backbone network, ICG, and its associated cities (Figure 10.1).

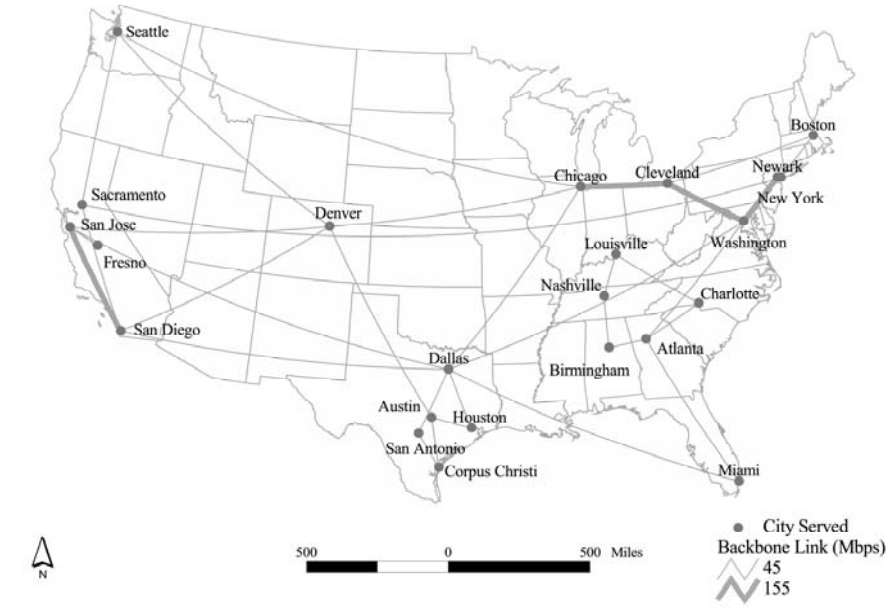


Fig. 10.1 ICG Network Backbone

As noted previously, the ICG network consists of 23 cities and 36 intercity linkages. Each link on the network is assigned bandwidth capacity measured in megabits per second (Mbps), serving as a_j in this analysis. Similarly, each city is assigned a nodal weight representing its total population for the year 2000. This measure serves as c_i for the analysis. The network system was evaluated using BO-NRIP through the use of Avenue scripts in ArcView to write the associated integer program to a text file. The problem is then solved externally to ArcView using CPLEX ver. 9.0.

A result file is then exported from CPLEX and read into ArcView for subsequent display and
analysis.

Node Removal: Minimum Damage Scenario (Max – Min)

Using BO-NRIP, the max-min scenario reflects maximizing objective (1), Z_1, and minimizing objective (2), Z_2. As a weighted objective this takes the following form:

$$Maximize\ wZ_1 - (1-w)Z_2 \tag{8}$$

As structured, maximizing a negative component (Z_2 in this case), is the same as minimizing a non-negative component, assuming that $w \in [0,1]$. Therefore, the *max-min* version of BO-NRIP is a minimum damage scenario, representing the combination of node and arc losses where both a minimal amount of network bandwidth and nodal population is disrupted in the system. Table 10.2 displays solutions for this particular scenario. For $p = 1$, the results are not surprising. In the first instance, the total arc capacity objective is assigned an importance weight (w) of 1, so Fresno is the city identified by BO-NRIP as being best for removal. As noted previously, there is only one link that connects Fresno to the remainder of the ICG backbone network, a DS-3 (45 Mbps) backbone running to San Jose. Therefore, the loss of the Fresno node, and its incident link, maximizes the remaining bandwidth in the ICG system (i.e. minimizing damage). If we continue the examination of $p = 1$ with a slight adjustment to the weight corresponding to the arc capacity objective (w), the results vary. For example, when $w = 0.9$ (and $1\text{-}w = 0.1$), Birmingham is identified for removal. This modest change to the solution hints to the complex nature of this problem. Unlike Fresno, Birmingham maintains two unique DS-3 connections (one to Nashville and one to Atlanta). However, the population of Birmingham is almost 185,000 less than Fresno. Therefore, the slight shift in objective importance weighting to minimize the population impacted, changes the best solution. For alternative values of p, a similar trend

Table 10.2 ICG Node Removal (Max, Min)

p	Arc Weigh	Node Weigh	Nodes Removed	Aggregate Bandwidth	Population Impacted	Itera-tions	Time
1	1	0	Fresno	2,125	427,652	23	0.03
1	0.9	0.1	Birmingham	2,080	242,820	3	0.02
2	1	0	Fresno, Corpus Christi	2,035	705,106	79	0.08
2	0.9999	0.0001	Fresno, Birmingham	2,035	670,472	47	0.06
2	0.99	0.01	Louisville, Birmingham [Nashville]	1,945	499,051	6	0.02
3	1	0	Fresno, Nashville, Birmingham	1,990	1,215,996	71	0.06
3	0.99	0.01	Louisville, Birmingham, Corpus Christi	1,855	776,505	9	0.02
3	0.1	0.9	Newark, Louisville, Birmingham [Nashville]	1,590	772,597	8	0.01
4	1	0	Fresno, Austin, Houston, Corpus Christi	1,900	3,315,298	103	0.06
4	0.9999	0.0001	Louisville, Fresno, Nashville, Birmingham	1,900	1,472,227	50	0.06
4	0.99	0.01	Newark, Louisville, Birmingham, Corpus Christi [Nashville]	1,500	1,050,051	8	0
5	1	0	Fresno, Austin, Houston, San Antonio, Corpus Christi	1,855	4,459,948	66	0.05
5	0.9999	0.0001	Louisville, Fresno, Nashville, Charlotte, Birmingham	1,855	2,013,055	54	0.05
5	0.99	0.01	Newark, Louisville, Birmingham, Corpus Christi, Miami [Nash-ville]	1,410	1,412,521	12	0
6	1	0	Louisville, Fresno, Nashville, Charlotte, Atlanta, Birmingham	1,765	2,429,529	127	0.09
6	0.9999	0.0001	Louisville, Fresno, Nashville, Charlotte, Birmingham, Corpus Christi	1,765	2,290,509	51	0.06
6	0.99	0.01	Newark, Sacramento, Louisville, Birmingham, Corpus Christi, Miami [Nashville]	1,275	1,819,539	17	0.05

[City Subject to a Cascading Failure]*
*Not included in the population impacted calculation

is evident. For example, when $p = 5$ and $w = 1.0$ Fresno, Austin, Houston, San Antonio and Corpus Christi are identified as causing the least impact. In this case, the population impacted is 4,459,948. However, when the weights are adjusted to reflect more importance on population ($w = 0.99$), Newark, Louisville, Birmingham, Corpus Christi and Miami are identified. As one might expect, the population impacted is the $p = 5$ case is nearly 68% (1,412,521) less than the initial solution where $w = 1.0$.

Node Removal: Mixed Damage Scenarios

The mixed damage scenarios for BO-NRIP can take two forms. In the first form, one can seek to remove the nodes which maximize the overall amount of bandwidth available to the network while simultaneously maximizing the population impacted (*max-max*). The second mixed damage scenario for BO-NRIP minimizes the overall bandwidth available to the ICG backbone system while also minimizing the population impacted (*min-min*). Results for both cases are now discussed in turn.

(Max – Max)

The *max-max* scenario using BO-NRIP reflects maximizing objective (1), Z_1, and maximizing objective (2), Z_2. As a weighted objective this takes the following form:

$$Maximize\ wZ_1 + (1-w)Z_2 \tag{9}$$

Table 10.3 highlights the results of this case. For $p = 1$, where emphasis is placed on bandwidth ($w = 1$), Fresno is the obvious choice once again.

Table 10.3 ICG Node Removal (Max, Max)

p	Arc Weigh	Node Weigh	Nodes Removed	Aggregate Bandwidth	Population Impacted	Itera-tions	Time
1	1	0	Fresno	2,125	427,652	23	0.03
1	0.9999	0.0001	New York	1,970	8,008,280	1	0.02
2	1	0	Fresno, Corpus Christi	2,035	705,104	79	0.09
2	0.75	0.25	New York, Chicago	1,680	10,904,298	4	0.02
3	1	0	Fresno, Nashville, Birmingham	1,990	1,215,993	71	0.08
3	0.5	0.5	Chicago, New York, Houston	1,590	12,857,930	9	0
4	1	0	Fresno, Austin, Houston, Corpus Christi	1,900	3,315,298	103	0.11
4	0.9999	0.0001	Chicago, New York, Houston, San Antonio	1,500	14,002,580	53	0.05
4	0.98	0.02	Chicago, New York, San Diego, Houston	1,300	14,081,330	5	0.02
5	1	0	Fresno, Austin, Houston, San Antonio, Corpus Christi	1,855	4,459,945	66	0.08
5	0.9999	0.0001	New York, Austin, Houston, San Antonio, Corpus Christi	1,700	12,040,576	71	0.08
5	0.98	0.02	Chicago, New York, San Diego, Dallas, Houston	1,120	15,269,910	9	0.01
6	1	0	Louisville, Fresno, Nashville, Charlotte, Atlanta, Birmingham	1,765	2,429,529	127	0.09
6	0.99999	0.00001	Boston, Fresno, Austin, Houston, San Antonio, Corpus Christi	1,765	5,049,089	114	0.11
6	0.09	0.91	Chicago, New York, San Diego, Dallas, Houston, San Antonio	1,075	16,414,560	11	0.02

However, as the objective weight is adjusted to place more emphasis on population, the largest cities become better candidates for removal. For example, Table 10.3 illustrates that virtually any weighting scenario for $p = 1$ shifting emphasis to population (e.g. $w = 0.9999$), will encourage the selection of New York City (pop. 8,008,278). Similarly, while $p = 2$ solutions initially identify Fresno and Corpus Christi, slight variations in emphasis yield solutions with New York City and Chicago. However, the same cannot be said for higher values of p. Table 10.3 shows that for $p = 5$ there are three unique solutions to the problem. When emphasis is placed on bandwidth ($w = 1$), Fresno, Austin, Houston, San Antonio and Corpus Christi are identified for removal. This solution maximizes the remaining bandwidth in the system and minimizes the overall damage to the network. In other words, the backbone segments connecting these cities to the remainder of the network are the least important where aggregate system bandwidth are concerned. Nevertheless, a slight shift in weights to place additional emphasis on population ($w = 0.9999$) yields a different solution set, prompting the removal of New York, Austin, Houston, San Antonio

and Corpus Christi. In this instance, BO-NRIP is able to detect the "value" of removing New York as opposed to Fresno. Also of interest is the removal of Corpus Christi versus Fresno in this scenario. At face value, this seems to be an odd selection. The total population of Corpus Christi (277,454) is significantly less than Fresno (427,652). Corpus Christi is also connected to the ICG network with two DS-3 links. Therefore, in a solution where one is seeking to maximize bandwidth and maximize population impacted – why Corpus Christi? There are two reasons. First, at higher values of p, it is extremely difficult to compare individual values between cities (i.e. population or bandwidth). BO-NRIP seeks the best *combination* of cities/nodes for removal. As a result, by removing 80% of the nodes in Texas (including Corpus Christi), both population impacted and bandwidth are maximized. Second, by removing both Austin and San Antonio, their incident links are immediately deactivated. Because these are the only connections Corpus Christi maintains to the ICG network, the removal of this node and its 277,454 residents is a virtual "free-rider" because there is no additional bandwidth loss.

(Min – Min)

The min-min scenario for BO-NRIP reflects minimizing objective (1), Z_1, and minimizing objective (2), Z_2. As a weighted objective this takes the following form:

$$Minimize \;\; wZ_1 + (1-w)Z_2 \tag{10}$$

Results for this case are displayed in Table 10.4. The $p = 1$ solutions are good illustrations of this particular scenario.

Table 10.4 ICG Node Removal (Min, Min)

p	Arc Weigh	Node Weigh	Nodes Removed	Aggregate Bandwidth	Population Impacted	Itera-tions	Time
1	1	0	Washington	1,725	572,059	41	0.02
1	0.99	0.01	Birmingham	2,080	242,820	37	0.02
2	1	0	Cleveland, Newark	1,415	751,949	45	0.03
2	0.995	0.005	Newark, Birmingham	1,725	516,366	37	0.02
2	0.98	0.02	Louisville, Birmingham [Nasvhille]	1,945	499,051	33	0.01
3	1	0	Cleveland, Newark, Dallas	1,100	1,940,529	43	0.02
3	0.99	0.01	Newark, Louisville, Birmingham [Nashville]	1,590	772,597	30	0
4	1	0	Cleveland, Newark, San Jose, Dallas [Fresno]	855	2,835,472	48	0.03
4	0.99	0.01	Newark, Louisville, Birmingham, Corpus Christi [Nashville]	1,500	1,050,051	27	0.03
5	1	0	Cleveland, Newark, Denver, San Jose, Dallas [Fresno, Texas]	675	3,390,108	50	0.03
5	0.9999	0.0001	Cleveland, Newark, Denver, Atlanta, Dallas [Texas]	695	2,911,639	46	0.03
5	0.98	0.02	Newark, Louisville, Birmingham, Corpus Christi, Miami [Nash-ville]	1,410	1,412,521	25	0.03
6	1	0	Sacramento, San Jose, Fresno, Nashville, Charlotte, Atlanta	1,765	2,429,529	127	0.11
6	0.99999	0.0001	Cleveland, Newark, Denver, San Jose, Atlanta, Dallas [Fresno, Texas]				
6	0.099	0.901	Newark, Sacramento, Louisville, Birmingham, Corpus Christi, Miami [Nashville]	1,275	1,819,539	22	0.02

[City Subject to a Cascading Failure]* "Texas" = Houston, Austin, San Antonio and Corpus Christi
*Not included in the population impacted calculation

For example, because one is seeking to minimize overall bandwidth in the system, the selection of Washington ($w = 1$) in the first instance is not surprising. Washington maintains five unique connections in the system (Boston, Atlanta, Newark, Cleveland and Dallas). These connections represent the highest aggregate bandwidth total for any city at 445 Mbps. Once additional emphasis is placed on minimizing the population impacted, Washington becomes a less desirable target and Birmingham is identified. As noted previously, in addition to having two DS-3 links, Birmingham is the smallest city on the ICG backbone system.

For $p = 2$ several interesting results are discovered. When emphasis is placed on the arc bandwidth objective ($w = 1$), the selection of Cleveland and Newark for removal is important, as noted in Grubesic and Murray (2006). First, the highest bandwidth city, Washington, is not included in the solution set. Second, in terms of population, Cleveland and Newark are much smaller than other cities included on the network. Both of these facts hint to the relatively non-intuitive nature of launching an optimal targeted attack on any network, particularly if the focus is on bandwidth reduction. This solution also suggests that the highest profile cities, such as Chicago, New York and Los Angeles, do not always represent the best locations for a coordinated attack. Instead, the most important locations in a backbone system are the vital nodes that maintain the majority of system interconnections (Grubesic and Murray, 2006). Although these locations are often positively correlated to major public peering points (e.g. NAPs and MAEs), in cities like Chicago, New York and Washington, it is not always the case (see Grubesic and O'Kelly 2002). This type of network topology, where a select set of nodes maintains the majority of connections, is often described as "scale free" (Barabasi et al. 2000; Comellas et

al. 2000; Barabasi et al. 2001; ben-Avraham et al. 2003). Unfortunately, these types of topological structures are quite vulnerable to failure if vital nodes are attacked (Albert et al. 2000).

One of the more interesting scenarios occurs with the removal of three nodes in the ICG backbone system. When the node attribute objective is given additional emphasis (1-w = 0.01), Louisville, Birmingham and Newark are selected for removal.

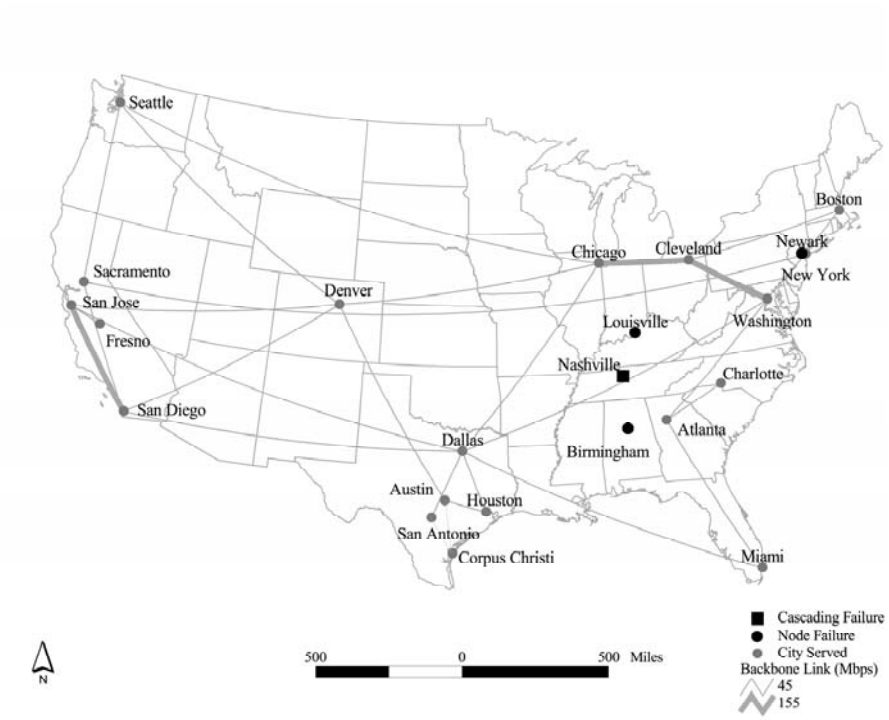


Fig. 10.2 ICG Node Removal, p = 3 (Min, Min)

However, instead of simply losing three cities, the loss of Louisville and Birmingham prompts a cascading failure in the ICG network, disconnecting Nashville (Figure 10.2). Similarly, when four nodes are selected for removal and priority is given to the arc attribute (w = 1), Cleveland, Newark, San Jose and Dallas are flagged as the best candidates for removal - but the loss of San Jose also prompts a cascading failure in the ICG network, disconnecting Fresno (Figure 10.3).

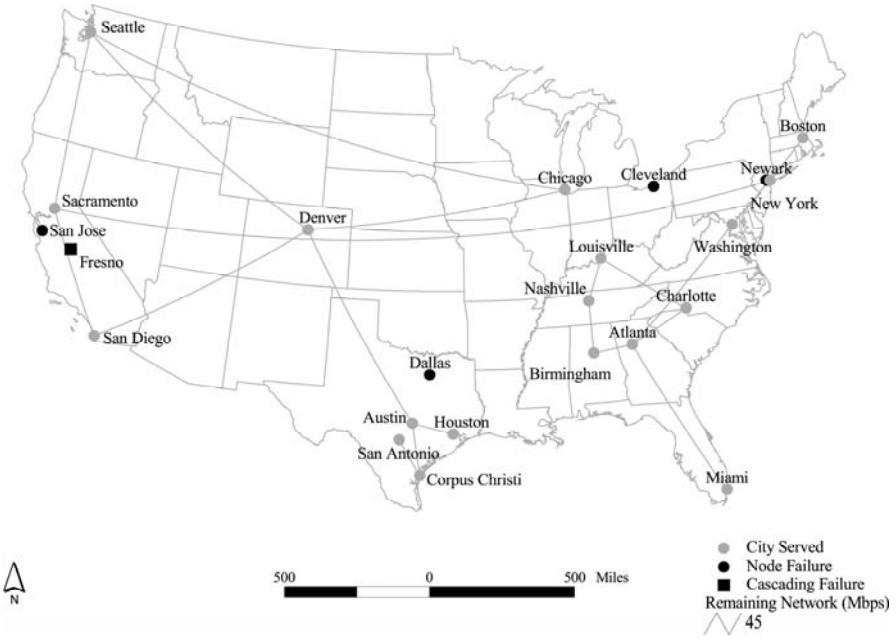


Fig. 10.3 ICG Node Removal, p=4 (Min, Min)

At the very least, this suggests that the loss of three or four nodes leaves the ICG network in a state of relative fragility that is not evident for lower values of p.

Node Removal: Maximum Damage Scenario (Min-Max)

The final scenario, min-max, for BO-NRIP reflects minimizing objective (1), Z_1, and maximizing objective (2), Z_2. As a weighted objective this takes the following form:

$$Minimize \; wZ_1 - (1-w)Z_2 \tag{11}$$

This scenario seeks to minimize the aggregate bandwidth available and maximize the population impacted. The solutions for this scenario are presented in Table 10.5. It is important to note that the *min-max* scenario mirrors solutions in the *min-min* scenario when the emphasis is placed on arc weight (w = 1). Therefore, the most interesting solutions for *min-max* occur when weights are shifted to increase the importance of population in the objective functions. For example, if we are strictly concerned with

bandwidth, at $p = 5$, Cleveland, Newark, Denver, San Jose and Dallas are selected for removal. As noted previously, this creates a significant problem for the network, with the associated cascading failures disconnecting many more than five cities (e.g. Fresno and all the remaining nodes in Texas). A slight shift in weight ($w = 0.9999$) changes the solution to include Chicago, New York, Washington, San Diego and Dallas. In this instance, there is an interesting dynamic between bandwidth and population. Although the weighting scheme places the majority of the emphasis on bandwidth, the influence of population is notable, particularly with the entry of Chicago, New York and San Diego into the solution. However, the entry of Washington is of great interest. Although the population of Washington is lower than other cities on the ICG network (e.g. Boston, Austin or San Antonio), the presence of two OC-3 (155 Mbps) connections prompted its inclusion. As noted previously, these OC-3 connections contribute to Washington's status as the highest bandwidth city on the ICG network. Additional shifts in the weighting scheme allow for the true maximization of population impacted. For example, when ($w = 0.1$), BONRIP identifies the five largest cities (Chicago, New York, San Diego, Dallas, Houston and San Antonio) for removal, impacting nearly 16.5 million people and over 50% of the aggregate system bandwidth.

Table 10.5 ICG Node Removal (Min, Max)

p	Arc Weigh	Node Weigh	Nodes Removed	Aggregate Bandwidth	Population Impacted	Itera-tions	Time
1	1	0	Washington	1,725	572,059	41	0.02
1	0.9999	0.0001	New York	1,970	8,008,280	34	0.02
2	1	0	Cleveland, Newark	1,415	751,949	45	0.02
2	0.99999	0.00001	Chicago, Washington	1,435	3,468,079	44	0.05
2	0.9999	0.0001	Chicago, New York	1,680	10,904,300	36	0.02
3	1	0	Cleveland, Newark, Dallas	1,100	1,940,529	43	0
3	0.9	0.1	Chicago, New York, Houston	1,590	12,857,930	29	0.02
4	1	0	Cleveland, Newark, San Jose, Dallas [Fresno]	855	2,835,472	48	0.02
4	0.9999	0.0001	Chicago, New York, Washington, San Diego	945	12,699,759	35	0.02
4	0.9	0.1	Chicago, New York, San Diego, Houston	1,300	14,081,330	25	0.02
5	1	0	Cleveland, Newark, Denver, San Jose, Dallas [Fresno, Texas]	675	3,390,108	50	0.03
5	0.9999	0.0001	Chicago, New York, Washington, San Diego, Dallas	765	13,888,339	33	0.02
5	0.9	0.1	Chicago, New York, San Diego, Dallas, Houston	1,120	15,269,910	21	0.02
6	1	0	Cleveland, Newark, Denver, Altanta, San Diego, Dallas [Texas]	495	4,135,039	52	0.02
6	0.9999	0.0001	Chicago, New York, Washington, San Diego, Dallas, Houston	720	15,841,969	31	0.05
6	0.99	0.01	Chicago, New York, San Diego, Dallas, Houston, San Antonio	1,075	16,414,560	20	0.02

[City Subject to a Cascading Failure]* "Texas" = Houston, Austin, San Antonio and Corpus Christi
*Not included in the population impacted calculation

10.5 Discussion and Conclusion

Given the increasing political, social and economic reliance on geographically linked networks (e.g. telecommunication, electrical, water, etc.), the notion of continuity is more important than ever. The empirical results presented in this chapter suggest several trends worth further discussion. First, continuity in critical network infrastructure is highly contingent upon topology. Overall levels of system vulnerability dramatically increase when critical network nodes lack diversity and redundancy (Grubesic and Murray 2005; Grubesic et al. 2003; White House 2003). Perhaps the best example of this is the removal of a specific node that prompts the cascading failure of additional cities/nodes which do not maintain secondary or tertiary connections to the network. In the case of the ICG network, the loss of San Jose and its incident arcs also prompts the loss of Fresno. However, this is not to say that nodes which maintain two or more connections are not vulnerable. As demonstrated in Figure 10.3, the loss of Birmingham and Louisville also prompted the cascading failure of Nashville, a city which had maintained two connections to the ICG network. At the very least, the results presented in this chapter corroborate previous studies which suggest that both hub-and-spoke and scale-free networks are extremely vulnerable to attack, particularly if assets are concentrated at a select set of vital nodes (Grubesic and Murray 2005; Barabasi et al. 2000; Comellas et al. 2000; Barabasi et al. 2001; ben-Avraham et al. 2003).

A second point worth noting centers on the applicability of BO-NRIP. As noted previously, both of these optimization models are flexible enough to be applied to any geographically linked network. For the purposes of this chapter, BO-NRIP was used to model optimal targets for either minimizing or maximizing network damage by removing nodes in a telecommunication system. However, there is the potential to extend the applicability of this family of models to network recovery and disaster mitigation. For example, if the ICG network suffered the loss of n nodes due to a natural disaster or terrorist attack, the first priority would be to mitigate this damage and begin the network recovery process immediately. Assuming that ICG has limited human and network resources for disaster recovery, mitigation requires some level of prioritization during the restoration process. For instance, one goal might be to maximize the restoration effort based on population or subscribers. A different goal might be to maximize the restoration effort based on bandwidth. A third goal might include a combination of population and bandwidth. Given any of these scenarios, it would be worthwhile to explore the characteristics of a node insertion approach for modeling an optimal network recovery plan. That is, instead of

removing nodes, one reactivates damaged nodes and their incident arcs in an optimal pattern.

Finally, there is a strong need to nurture the development of coordinated federal, state and local policies to offset the dangers associated with interdependencies between geographically linked systems. As noted in Section 10.1, the federal government is keenly aware of the problems associated with asset concentration, isolated interconnections between systems, and the potential economic losses that a major disruption could cause. However, the vast majority of critical infrastructure in the United States is privately held. As a result, the U.S. government is hoping that their latest efforts to promote data-sharing and emergency preparation through the Department of Homeland Security can provide a solid foundation for coordinating national, state and local efforts in securing critical infrastructure.

This chapter examined the spatial ramifications of removing nodes in a geographically linked network. In addition, the modeling framework presented in this chapter provided an extension to previous work by Grubesic and Murray (2006), addressing nodal attributes explicitly. This bi-objective model helps incorporate a new level of geographic context to critical infrastructure analysis by providing an estimate of potential damage during a major disaster. Finally, this chapter serves to answer the call for more work in the areas of critical infrastructure protection and modeling, a subfield of emerging importance to the spatial and economic planning sciences (Cutter et al. 2003; Church et al. 2004; Grubesic and Murray 2006).

References

Albert, R., I. Albert, and G. L. Nakarado. 2004. *Physical Review E*. 69:(025103) 1-4.

Albert, R., H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature*. 406: 378 – 382.

Barabasi, A. L., R. Albert, and H. Jeong. 2000. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*. 281: 69-77.

Barabasi, A. L., E. Ravasz, and T. Vicsek. 2001. Deterministic scale-free networks. *Physica A*. 299: 559-564.

ben-Avraham, D., A. F. Rozenfeld, R. Cohen, and S. Havlin. 2003. Geographical embedding of scale-free networks. *Physica A*. 330: 107-116.

Botelho, G. 2003. Power returns to most areas hit by blackout. *CNN*. URL: http://www.cnn.com/2003/US/08/15/power.outage/

Boedhihartono, P. and G. Maral. 2003. Handover evaluation in non-geostationary satellite constellation systems implementing mutual visibility and diversity. *Space Communications*. 19(1): 23-45.

Carreras, B. A., V. E. Lynch, I. Dobson, and D. E. Newman. 2002. Critical points and transitions in an electric power transmission model for cascading failure blackouts. *Chaos*. 12(4): 985 – 994.

Chang, S. E., W. D. Svekia, and M. Shinozuka. 2002. Linking infrastructure and urban economy: simulation of water disruption impacts in earthquakes. *Environment and Planning B*: *Planning and Design*. 29: 281-301.

Church, R. L., M. P. Scaparra, and R. S. Middleton. 2004. Identifying Critical Infrastructure: The Median and Covering Facility Interdiction Problems. *Annals of the Association of American Geographers*. 94(3): 491-502.

Church, R. L. and M. P. Scaparra. 2005. Protecting critical assets: The *r*-interdiction median problem with fortification. Working Paper No. 79. Kent Business School.

Cohan, J. 1978. *Multi-objective programming and planning*. Academic Press, New York.

Colbourn, C. J. 1999. Reliability issues in telecommunications network planning'. In, *Telecommunications Network Planning*, ed. Sans, B. Kluwer Academic Press.

Comellas, F., J. Ozon, and J. G. Peters. 2000. Deterministic small-world communication networks. *Information Processing Letters*. 76, 83-90.

Coutard, O, Hanley, R.E., and R. Zimmerman. 2005 (eds.), *Sustaining urban networks: The social diffusion of large technical systems.* New York: Routledge.

_____. 2005. Network systems revisited: the confounding nature of universal systems. In, *Sustaining urban networks: The social diffusion of large technical systems*, eds. Coutard, O., Hanley, R.E. and R. Zimmerman. New York: Routledge.

Corley, H. W. and H. Chang. 1974. Finding the n most vital nodes in a flow network. *Management Science*. 21: 362-364.

Cutter, S. L., D. B. Richardson, and T. J. Wilbanks. 2003. A research and action agenda. In *The geographical dimensions of terrorism*, ed. S. L. Cutter, D. B. Richardson, and T. J. Wilbanks, 223-29. New York: Routledge.

Faisal, H. 2005. Task to detect major fault in Pakistan's Internet cable set off. *Pakistan Times*.
URL: http://pakistantimes.net/2005/07/04/top1.htm

Grubesic, T.H., O_Kelly, M.E., 2002. Using Points of Presence to Measure City Accessibility to the Commercial Internet. *Professional Geographer*. 54 (2): 259–278.

Grubesic, T. H. and A. T. Murray. 2005. Spatial-historical landscapes of telecommunication survivability. *Telecommunications Policy*. 29(11): 801 – 820.

Grubesic, T. H. and A. T. Murray. 2006. Vital nodes, interconnected infrastructures and the geographies of cascading failure. *Annals of the Association of American Geographers*. 96(1): 64-83.

Grubesic, T. H., M. E. O'Kelly, and A. T. Murray. 2003. A Geographic Perspective on Commercial Internet Survivability. *Telematics and Informatics*. 20: 51-69.

Houck, D. J., E. Kim, G. P. O'Reilly, D. D. Picklesimer, and H. Uzunalioglu. 2004. A network survivability model for critical national infrastructures. *Bell Labs Technical Journal*. 8(4): 153-172.

Johnson, J. and A. Lefebvre. 2003. U.S. Impact of the Northeast Blackout continues to emerge. URL: http://www.wsws.org/articles/2003/aug2003/blck-a20.shtml

Jrad, A., Morawski, T. and L. Spergel. 2004. A model for quantifying business continuity preparedness risks for telecommunications networks. *Bell Labs Technical Journal*. 9(2): 107-123.

Kansal, M. L., Kumar, A. and P. B. Sharma. 1995. Reliability analysis of water distribution systems under uncertainty. *Reliability Engineering & System Safety*. 50: 51-59.

Little, R. G. 2002. Controlling Cascading Failure: Understanding the Vulnerabilities of Interconnected Infrastructure. *Journal of Urban Technology*. 9(1): 109-123.

Medhi, D., 1999. Network reliability and fault-tolerance. In, *Encyclopedia of Electrical and Electronics Engineering*, ed. Webster, J. Wiley. John Wiley and Sons, New York.

Mohamed, A.A., Leemis, L. M. and A. Ravindran. 1992. Optimization techniques for system reliability – A review. *Reliability Engineering & System Safety*. 35: 137-146.

O'Kelly, M.E. and M. W. Horner. 2003. Aggregate Accessibility to Population at the County Level, U.S. 1940-2000. *Journal of Geographical Systems*, 5(1): 5-23.

Premkumar, G., Chou, C. H. and H. H. Chou. 2000. Telecommunications network design – Comparison of alternative approaches. *Decision Sciences*. 31: 483-506.

Ratliff, H. D., Sicilia, G. T. and S. H. Lubore. 1975. Finding the n most vital links in flow networks. *Management Science*, 21(5): 531-539.

Reuters. 2005. Internet crashes in Pakistan. *CNN*. URL :http://www.cnn.com/2005/WORLD/asiapcf/06/28/pakistan.internet.reut/index.html

Talukdar, S. N., J. Apt, M. Ilic, L. B. Lave, and M. G. Morgan. 2003. Cascading Failures: Survival versus Prevention. *The Electricity Journal*. November: 25-31.

Tillman, F. A., Hwang, C. L., and W. Kuo. 1977. Determining component reliability and redundancy for optimum system reliability.

U.S. Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunications Services. (1996). *Federal Standard 1037C*.

Wollmer, R. 1964. Removing arcs from a network. *Operations Research*. 12(6): 934-940.

White House. (2003). The National Strategy for the Physical Protection of Critical Infrastructures and Key Assets.
URL: http://www.whitehouse.gov/pcipb/physical_strategy.pdf

Zimmerman, R. 2005. Social implications of infrastructure network interactions. In, *Sustaining urban networks: The social diffusion of large technical systems*, eds. Coutard, O., Hanley, R.E. and R. Zimmerman. New York: Routledge.

11 Analysis of Facility Systems' Reliability When Subject to Attack or a Natural Disaster

Richard Church[1] and M. Paola Scaparra[2]

[1] Department of GeographyUniversity of California, Santa Barbara Santa Barbara, California, USA ; Email: church@geog.ucsb.edu
[2] Kent Business School, University of Kent; Canterbury, Kent, UK; Email: M.P.scaparra@kent.ac.uk

11.1 Introduction

Critical infrastructure can be defined as those elements which are necessary for lifeline support and safety. They include such systems as communication systems, water and sewer systems, health services facilities, food production/processing/storage systems, transportation systems, drug production/stockpiles, and incident sensing/detection/control systems. Each of these systems has unique properties that may define specific issues in operation and management in order to provide a consistent and continuing level of operation. A common question today is whether a particular system or component is vulnerable to failure and whether in some cases failure of one system component will lead to a failure of downstream components. For example, the electrical system failure in Ohio led to significant loss of power in many states of the Northeast US in August, 2003. Grubesic, et al. (2003) has called this cascading failure. There are five major problems in managing, operating and designing infrastructure: 1) for existing systems, identify those components that are subject to natural disasters along with their impacts on system operation; 2) for existing systems, identify those components that if chosen by an interdictor, would impact system operations the most; 3) for existing systems, identify those systems components that should be protected against natural or intentional strikes, in order to keep a system operating efficiently; 4) for a new system, design a system so that its operation is as resilient as possible against natural disasters and intentional strikes; and 5) schedule and allocate protection re-

sources in order to minimize disruptions and impacts on system efficiency due to natural losses or intentional strikes. These five problem areas capture a range of issues associated with keeping a "lifeline" system in operation.

In this chapter we address the issues raised in problem areas 1 and 2. We assume that we have a somewhat generic system in operation and that it contains a set of manufacturing/processing plants. Each of these facilities serves a customer or demand set, whereby customers are supplied by its closest facility in terms of cost (or time) of service. Think of the product or services as important. Services must be provided if at all possible or as long as at least one facility exists. If a facility is compromised or lost, the system will operate at an impaired level of operation, with attendant additional costs of transportation and production. At issue is the impact of system efficiency when one or more facilities are lost due to a natural disaster or intentional strike. The problem that we address was first proposed by Church, Scaparra, and Middleton (2004). The paper by Church et al. also presents a review of the literature associated with network interdiction, so we will not review the literature here. In Church et al., the objective was to identify the set of facilities which if lost, would impact system operation efficiencies the most. Two different systems were modeled, an emergency service delivery system, and a production/transportation system. In this chapter we will concentrate our analysis on a generic production/delivery system.

Church et al. (2004) presented a model called the *r*-interdiction median problem. This model can be used to identify which *r* of the existing set of *p* facilities, when interdicted or lost impacts delivery efficiency the most. Such a model can be used to identify the worst-case of loss, when losing a pre-specified number of facilities. The model is restricted in two ways: it is based upon the assumption that the terrorist or interdictor is successful in each and every strike, and it is also based upon the assumption that exactly *r* facilities will be struck and lost. Such a model does address a worst-case scenario, but it does not exactly capture the issues that would be key to understanding the range of failures and possible outcomes. First, it is important to recognize that a strike or disaster may not impair a facility's operation. That is, a terrorist strike may be successful only a certain percentage of the time. The same is true for a natural disaster. When it does occur, there is a threat that operations at a facility may need to be suspended, but it is not an absolute. Second, interdiction may not be intelligent whereby the strike is to a facility that is not so critical to overall system operations. Although it is important to model "worst-case" scenarios, it is also important to model and understand the range of possible failures and impacts. The overall objective of this chapter is to present a family of models which

can be used to model the range of possible impacts associated with the threat of losing one or more facilities to a natural disaster or intentional strike. We show how to model two cases, deterministic loss and probabilistic loss. In addition, we present results associated with the application of worst-case and best-case expected loss models to a data set used previously in modeling system impacts associated with production/supply systems. The modeling framework that we present is new and innovative, and can be expanded to other systems as well.

In the next section we discuss the *r*-interdiction median problem of Church, et al. (2004) as well as the concept of a reliability envelope. We also discuss how the *r*-interdiction median problem can be used with another model to generate a deterministic reliability envelope. Following that, we present two new models to address probabilistic losses due to natural or man-inflicted disasters. Then, we present results from both sets of models and follow with a summary.

11.2 The Reliability Envelope

Suppose that we have a system of *p* operating facilities supplying a set of *n* demand points. If each facility can serve any assigned demand, then we can assign each customer to their closest facility (as measured by cost or distance). We can define weighted distance for a demand-facility interaction as the distance from the demand to their closest facility weighted by the number of trips needed to supply that demand from a facility utilizing some type of transport mode (e.g. truck). Thus, we can measure the overall efficiency of the system as the total truck-miles of travel needed to supply all of the demand from the set of located facilities. In location science, the problem of locating *p*- supply facilities that yields the smallest weighted distance is called the *p*-median problem. The *p*-median problem has been the subject of considerable research, starting with the theorems of Hakimi (1964, 1965), the first heuristic of Teitz and Bart (1968) and an integer linear-programming model of ReVelle and Swain (1970). Church (2003) provides a detailed summary of different approaches for solving the *p*-median problem.

The exact opposite of the *p*-median problem occurs when you have an existing system of *p* facilities, which may or may not be located optimally. Of these *p* facilities, we may lose some facilities (e.g. to a natural disaster or due to an intentional strike by a terrorist or interdictor). When either closing or considering the loss of one or more facilities by a disaster, the

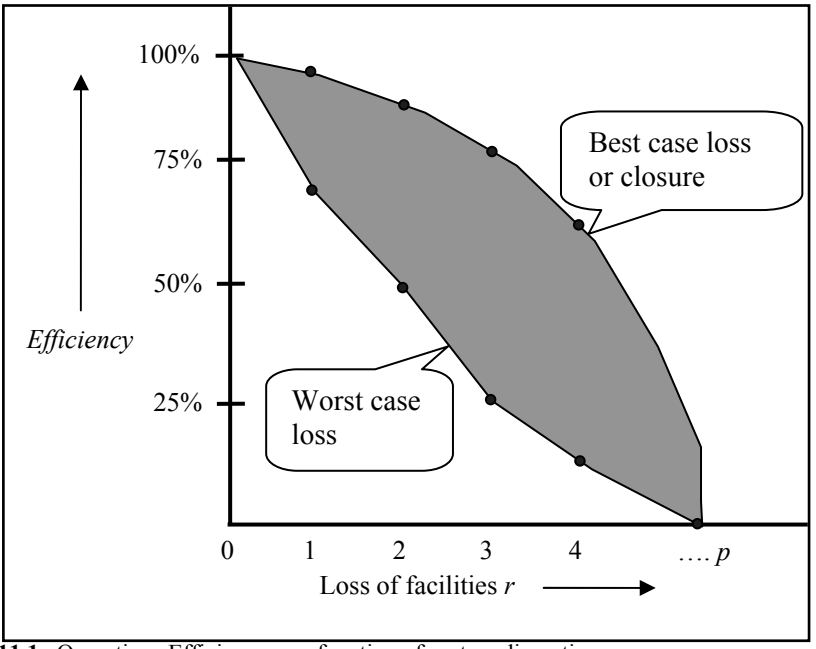


Fig. 11.1. Operations Efficiency as a function of system disruption

basic question is what happens to the operating efficiency of the remaining system. We can measure this loss of efficiency by calculating the resulting increase in weighted distance. We can depict the increase in weighted distance (or loss of system efficiency) as shown in Figure 11.1. The *x*-axis represents the level of facility losses or closures, associated with an existing configuration of *p* facilities. The *y*-axis depicts resulting system efficiency. For this example, efficiency is measured at 100% if all *p* facilities are operating. If a facility is lost due to a natural disaster, intentional strike, or planned closure then efficiency is lost and overall efficiency decreases. If all facilities are closed or lost, then we can say that the system is no longer in operation and has no level of efficiency. If many facilities exist, then there exist several possible outcomes of losing just one facility. One can easily enumerate each of the possible ways of losing one facility as well as calculate the impact of each possible loss in terms of changes in efficiency. The results of this series of calculations will define a range of losses from best-case (i.e. least decrease in efficiency) to worst-case (i.e. greatest decrease in efficiency). In Figure 11.1, the darkened region is defined by an upper curve and a lower curve. The upper curve represents the solutions of least impact associated with a given loss level. The lower curve represents the solutions of greatest impact associated with a given loss level. For our case of enumerating the impacts of all possible ways of

losing a facility, the best-case defines the point on the upper curve and the worst-case depicts the solution defining the point on the lower curve when r =1. The region that is depicted between these two curves can be defined as the operational envelope or reliability envelope. This type of diagram is similar to the type of envelope developed by Kim and O'Kelly (2004) in order to depict possible outcomes for the failure of a communication system (see also Baran (1964) and Urban and Keitt (2004) as other examples is depicting operational envelopes). Here we use the concept to depict the operational region of a set of facilities which may be subject to losses. Computing system reliability for networks with probabilistic rates of component and link failures began with the ground-breaking work of Moore and Shannon (1956) and Birmbaum et al. (1961). Much of the work on estimating the reliability of a given network is based upon two assumptions, *i.e.* events are random and independent. These two assumptions do not necessarily hold when considering terrorism, cascading failures, and large scale disasters (e.g. earthquake clusters).

The region contains many possible outcomes, defined by a minimum and a maximum efficiency level for each level of loss, r. Determining the exact nature of the envelope can be a valuable aid to decision-making. The upper curve represents a situation of complete facility control where some of the facilities need to be closed due to some limitation (for example a budget cut, or the loss of some level of raw material supply which necessitates the closure of one or more facilities). In this case the decision maker can decide exactly which facilities to close in order to minimize the impact of closures. The lower curve represents the opposite, where the loss of one or more facilities is not within the control regime of the decision-maker. This is represented by either a natural disaster or an attack by someone intending to disrupt the system. Knowing just how much a system can be impacted by such events can be valuable in planning contingencies, redesign, and fortification/protection. Another important property to understand is the nature of the thickness of the envelope. The difference between best and worst case outcomes for a value of r, help define the upper limits on possible improvements to system efficiency due to contingency planning. It is important to understand that the envelope can be determined through the use of two different models that have been developed in the literature. The first is the p-median model and the second is the r-interdiction median model. In the remainder of this section we will define these two models and describe how the operational envelope can be efficiently generated.

We begin the discussion of generating the envelope by concentrating on the upper curve, or the best/optimistic case loss. This occurs when we can choose which r of the p facilities to close. The best-case is to close those facilities which have the least impact on weighted distance or keep those

facilities which keep weighted distance as low as possible. We assume here that we have a network where each node is a point of demand, and a subset of p nodes which house facilities. We will formulate the model to choose the best facilities to close using the following notation:

i, j = indices used to refer to a node, numbered as 1,2,…,*n*.

d_{ij} = shortest distance from node *i* to node *j*

a_i = demand at node i

$x_{ij} = \begin{cases} 1, \text{ if demand at } i \text{ assigns to facility at } j \\ 0, \text{ otherwise} \end{cases}$

$y_j = \begin{cases} 1, \text{ if the facility at site } j \text{ is kept open} \\ 0, \text{ otherwise} \end{cases}$

p = the number of existing facilities

r = the number of existing facilities that are to be closed

E = the set of existing facility sites

Using the notation defined above, we can formulate an optimistic closing model as the following integer-linear programming problem:

$$Min\ Z = \sum_{i=1}^{n} \sum_{j \in E} a_i d_{ij} x_{ij} \tag{1}$$

Subject to

$$\sum_{j \in E} x_{ij} = 1 \text{ for each } i = 1, 2, ..., n \tag{2}$$

$$\sum_{j \in E} y_j = p - r \tag{3}$$

$$x_{ij} \leq y_j \text{ for each } i = 1,2,....n \text{ and } j \in E \tag{4}$$

$$x_{ij} = 0,1 \text{ for each } i = 1,2,....n \text{ and } j \in E \tag{5}$$

$$y_j = 0,1 \text{ for each } j \in E \tag{6}$$

The above model selects which *p*-*r* facilities to keep open in order to minimize the resulting weighted distance. By keeping *p*-*r* of *p* facilities, *r* of the facilities are closed or eliminated. Constraint (2) ensures that each demand must assign to a facility that remains open. Constraint (3) specifies that *p*-*r* facilities are kept open, thus *r* facilities are closed. Constraints of type (4) restrict a demand to assign only to a facility that is kept open. Finally, constraints (5) and (6) specify the integer restrictions of the model. It can be easily shown that constraints (5) are not necessary in order to generate solutions which are integer (although each variable must be restricted to be non-negative in value). This model decides which facilities to keep open so that the resulting weighted distance is minimized. It is important to recognize that this model is a special form of the classic *p*-median problem. In fact, it is a special case of the model where the site set is restricted to the set of existing sites, *E*. The *p*-median literature contains a number of different solution approaches; virtually every one of them can be applied to solve this particular model form (Church, 2003).

We now turn our attention to the generation of the bottom curve of the envelope, which represents the worst of the possible circumstances for losses. We can approach the problem of generating worst-case losses from the perspective of an indictor. An interdictor would select those facilities/sites which when removed from the system, yields the greatest increase in total weighted distance. This problem was first defined by Church, Scaparra and Middleton (2004) and was called the *r*-interdiction median problem. It is easy to see that the interdiction problem is the antithesis of the *p*-median problem. Whereas the *p*-median location problem involves locating *p* facilities in order to maximize efficiency, the *r*-interdiction median problem involves finding the best subset of existing supply sites to remove in order to minimize efficiency. We can formulate this problem using the following additional notation:

$$s_j = \begin{cases} 1, & \text{if a facility located at } j \text{ is eliminated, i.e. interdicted} \\ 0, & \text{otherwise} \end{cases}$$

r = the number of facilities to be interdicted or eliminated

$T_{ij} = \{k \in E \mid k \neq j \text{ and } d_{ik} > d_{ij}\}$, the set of existing sites (not including *j*) that are as far or farther than *j* is from demand *i*.

We can now formulate the r-interdiction median (RIM) problem as the following integer- linear programming problem:

$$Max \;\; Z = \sum_{i=1}^{n} \sum_{j \in E} a_i d_{ij} x_{ij} \tag{7}$$

Subject to:

$$\sum_{j \in E} x_{ij} = 1 \quad for\ each\ \ i = 1,2,....n \tag{8}$$

$$\sum_{j \in E} s_j = r \tag{9}$$

$$\sum_{k \in T_{ij}} x_{ik} \leq s_j \quad for\ each\ \ i = 1,2,....,n\ and\ each\ j \in E \tag{10}$$

$$\begin{aligned} x_{ij} &= 0,1 \quad for\ each\ i = 1,2,...,n\ \ and\ \ each\ j \in E \\ s_j &= 0,1 \quad for\ each\ j \in E \end{aligned} \tag{11}$$

The objective of this model (7) seeks to maximize the weighted distance upon the interdiction of *r* facilities. That is, the objective seeks the solution which has the greatest impact on weighted distance, when closing or removing *r* facilities from the existing configuration. Constraint (8) specifies that each demand must assign to a facility after interdiction/removal. Constraint (9) sets the number of interdicted facilities to equal *r*. Constraints (10) help ensure that a given demand *i* assigns to its closest remaining facility *j*. Constraints (10) are designed to prevent demand *i* from assigning to any facilities that are farther than *j* is to *i*, unless the facility at *j* has been subject to interdiction. Thus, demand *i* is forced to assign to its closest open facility after interdiction. Integer restrictions on the variables are specified in (11). It is important to note that the integer restrictions on the x_{ij} variables are not necessary when solving RIM by general purpose mixed-integer programming software. This means that only the s_j variables will need to be forced to be zero-one. Church, Scaparra and Middleton (2004) used general purpose optimization software to solve the RIM model for a set of example problems. In a subsequent work, Scaparra and Church (2006) show how the RIM formulation can be streamlined through a process of variable reduction and consolidation similar to the one proposed by Church (2003) for the *p*-median problem. The condensed formulation could be solved by a mixed-integer solver in significantly less time

than the original formulation. Scaparra and Church (2006) also demonstrate through empirical tests that constraints (10), initially proposed by Church and Cohon (1976), are the most efficient way of enforcing closest assignment in RIM models. RIM formulations with different closest assignment constraints, such as the ones proposed by Rojeski and ReVelle (1970), turn out to be more difficult to solve, especially for larger values of *r*.

The two models described above can be applied to a given system over a range of *r* values from 1, 2, 3,…., *p-1*. The results can then be used to draw the points of the two curves which bound the operational envelope. For a given level of loss, this envelope specifies the range of possible system performance from best-case to worst-case. Actual performance will fall within this range. Monte Carlo simulation can then be used to generate a histogram or frequency of solutions which fall within each range, for each value of *r*. This means that models exist to solve for the reliability envelope when a specified level of loss is certain.

The operation or reliability envelope as defined so far is based upon a deterministic analysis. Up to this point we have considered a loss level, *r*, as certain. That is, if a facility was closed or interdicted, it was a certainty and it would happen. Thus, the above analysis can be defined as a deterministically derived envelope. Although the results of this analysis can be informing, it is important to recognize that an attack by an interdictor may not be successful, just as a natural disaster may or may knock out one or more facilities. This case cannot be handled by the above analysis. In the next section we extend this modeling framework to handle probabilistic losses.

11.3 Probabilistic Reliability Envelopes

Up to this point in the chapter, we have modeled site loss as a certainty. We now turn our attention to the case where loss is not a certainty. We know that whether a flood hits or a terrorist strikes, the chances of losing a facility are based upon some probability. Existing location-interdiction models do not handle this case. In this section we extend the modeling framework so that we can derive a reliability envelope when the loss of one or more facilities hit by a disaster is not certain. The closest related work to what we propose here is the work of Snyder and Daskin (2005). They developed a *p*-median location model which sites facilities in order to minimize expected losses. Their model was called the Reliability *p*-median problem (RPMP). The RPMP seeks those locations that, given

random acts of failure (represented by a probability), will perform with the best expected value. Thus, the model attempts to site facilities so that expected failure is accounted for. Here we wish to derive the envelope of expected efficiencies associated with an existing system. To do this we need to identify both worst-case and best-case expected outcomes which cannot be generated with RPMP. We will handle both cases: worst-case expected loss and best-case expected loss by the development of two new model forms. We will begin by modeling worst-case expected loss.

Our notation here builds upon the notation given in the previous section. We still assume that a set of r facilities will be hit by an interdictor or disaster. The difference here is that an interdiction or hit by a disaster may or may not be successful. We concentrate on the case where an interdictor can hit each facility at most once and that r facilities will be hit simultaneously. Think of this as a coordinated attack on the system by a terrorist and that all resources of the attack are used at once (rather than staged out). We will assume that the success of an interdiction upon a facility can be described as a probability:

θ = probability of success of an attempted interdiction

We consider that the probability of success on the part of the interdictor to be the same, regardless of the facility which is hit. We know that the probability of success may vary, but this extended case can be derived directly from the model that we develop here.

Attempts to interdict facilities may or may not impact a given demand i. This is depicted in Figure 11.2. Suppose that the closest facility to demand i is being considered for an attempted interdiction. This is represented at decision point 1 in the figure. If the closest facility to i is not attacked then that facility will remain in operation (case A in figure) and will continue to be the closest facility. If the decision is made to attack the closest facility to i, then it is possible that the service to demand i is degraded. This case is depicted as probabilistic node B. Either the attempt is successful or it is not. If the attempt to destroy the facility is not successful then the closest facility remains the first closest facility (case C in figure). If the first closest facility to i is attacked and it was successful in knocking out the facility, then the service must then be taken up by the second closest or even further facility. If the first closest facility is attacked with success, then we reach decision point 2. At decision point 2, the outcome rests on whether the second closest facility to i has been attacked. For if the second closest facility has not be subject to attack then we end at case D, where the closest available facility to i is the second closest facility. If the decision was made to attack the second closest facility to i along with a

successful attempt on the first closest facility to i, then we arrive at case E. If the attempt was not successful, then the second closest facility is still operating and serves demand i (case F). However, if the interdiction attempt was successful, then we arrive at decision point 3. The tree can con-

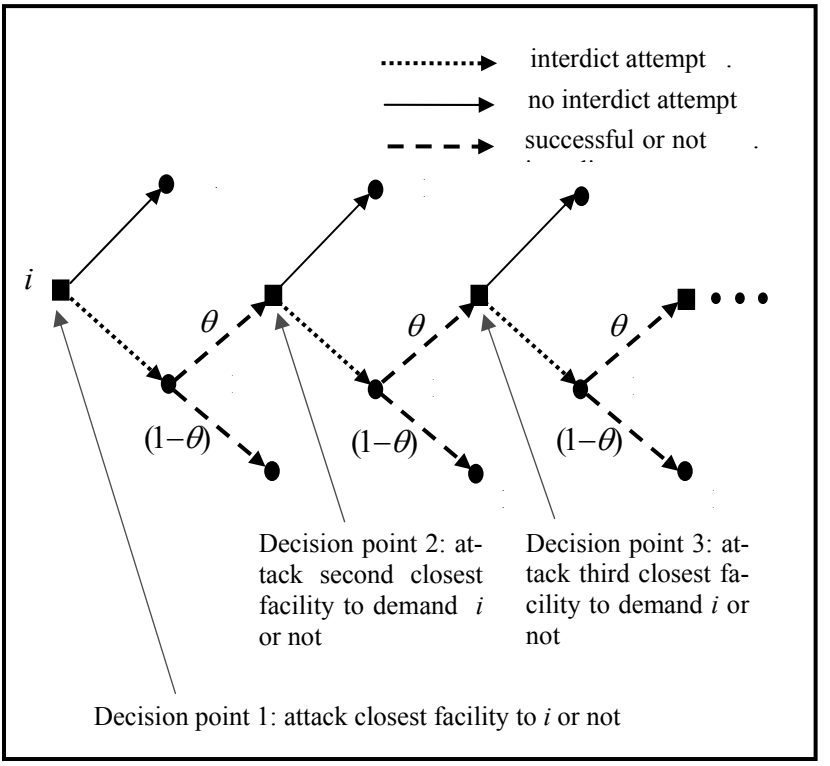


Fig. 11.2 Decision tree reflecting the states of possible interdiction cases inpamcting demand *i*. .

-tinue in this manner until we have considered the possible attack on all of the r closest facilities to node i. Even if all of the r closest facilities to i were attacked, it is possible that the first closest facility remains open (with probability $1-\theta$).

We can use the diagram in Figure 11.2 as a guide to calculate an expected distance to the closest open facility, should all of the k closest facilities to i be attacked and where an attack is successful with probability θ. This is:

$$d_i^k = \sum_{j=1}^{k} \theta^{j-1}(1-\theta) d_{ii_j} + \theta^k d_{ii_{k+1}} \tag{12}$$

where:

i_l = the index of the l^{th} closest facility to demand node i.

d_i^k = expected distance between demand i and its closest open facility given that the k closest facilities to i have been attacked with the probability that any given attempt is successful is θ

The impact of an attack on r out of p facilities can be calculated for demand i depending on whether the consecutive k closest facilities to i have been attacked. In fact, the tree and impact of a set of attacks for node i stops at the first decision node in which a decision not to attack is made. This is the key to building an interdiction model when success is not deterministic but probabilistic. To build such a model, we need to determine if the k closest facilities to a given demand i have all been attacked. We will do this through the use of the following decision variable:

$$T_{ik} = \begin{cases} 1, & \text{if the } k \text{ closest facilities to } i \text{ are attacked but the } k+1 \text{ closest is not} \\ 0, & \text{otherwise} \end{cases}$$

We also need to introduce a new decision variable to represent an attempt at interdiction as well as some additional notation:

$$\bar{s}_j = \begin{cases} 1, & \text{if facility } j \text{ is attacked} \\ 0, & \text{otherwise} \end{cases}$$

$F_k(i) = \{j \mid j \text{ is one of the } k \text{ closest sites to demand } i\}$

d_i^0 = distance between demand node i and its closest facility, which is the distance to serve i if the closest facility to i is not attacked

Using the above notation, we can formulate a model to optimize the attack on r out of p facilities so that the resulting expected weighted distance is maximized and where the success of any individual attack is given as θ. This model seeks to find the worst-case involving probabilistic outcomes of a set of r attacks. We call this model worst-case r-interdiction with probabilities median model (worst-case RIP). It is formulated as follows:

$$Max\ Z = \sum_{i=1}^{n}\sum_{k=1}^{r} a_i d_i^k T_{ik} + \sum_{i=1}^{n} a_i d_i^0 (1 - \bar{s}_{i_1}) \tag{13}$$

Subject to:

$$T_{ik} \le \bar{s}_l, \quad \text{for each } i = 1,2,...,n,\ k = 1,...,r, \text{ and each } l \in F_k(i) \tag{14}$$

$$\sum_{k=1}^{r} T_{ik} \le 1, \quad \text{for each } i = 1,2,...,n \tag{15}$$

$$\sum_{j \in E} \bar{s}_j = r \tag{16}$$

$$\bar{s}_j = 0,1 \text{ for each } j \in F \tag{17}$$

$$T_{ik} = 0,1 \quad \text{for each } i = 1,2,...,n \text{ and } k = 1,...,r \tag{18}$$

This model maximizes the expected weighted distance associated with a selected attack on *r* out of *p* facilities. If the closest facility to a given demand i is not attacked, then the expected distance of assignment is a constant, d_i^0. This is represented as the second term of the objective. If the closest facility to i is not attacked, then constraints of type (14) will keep all associated T_{ik} at zero in value. If the *k* closest facilities to i are attacked, for some value of *k*, then the second term of the objective will be zero and the first term will be positive. Constraints of type (14) will allow a given T_{ik} to be one as long as all *k* closest facilities are interdicted. Since constraint (15) prevents no more than one T_{ik} variable for a given demand to be one, then the model will select the T_{ik} with the highest d_i^k value. Thus, the model will identify the case where attacks occur for the highest *k* consecutive closest facilities to demand i. Constraint (16) limits the number of attacks to be equal to *r*, and the remaining constraints specify the integer restrictions on the decision variables. The "worst-case" RIP model can be used to generate a curve of expected weighted distance values for the lower portion of a probabilistic reliability envelope.

Now, we can turn our attention to modeling "best-case" expected loss. Here we seek to define the curve that defines the upper boundary of the probabilistic envelope. To do this we can build upon the notation developed for the worst-case model. In fact at issue is finding the set of r out of p facilities to attack that results in the lowest expected losses in efficiency or weighted distance. We can do this without introducing any additional notation as follows:

$$Min\ Z = \sum_{i \in N} \sum_{k=0}^{r} a_i d_i^k T_{ik} \tag{19}$$

$$T_{ii_k} \leq \bar{s}_{i_k}, \quad \text{for each } i = 1,2,...,n \text{ and } k = 1,2,...,r \tag{20}$$

$$T_{ii_k} \leq 1 - \bar{s}_{i_{k+1}}, \quad \text{for each } i = 1,2,...,n \text{ and } k = 0,...,r-1 \tag{21}$$

$$\sum_{k=0}^{r} T_{ik} = 1 \quad \text{for each } i = 1,2,...n \tag{22}$$

$$\sum_{j \in F} \bar{s}_j = r \tag{23}$$

$$\bar{s}_j = 0,1 \quad \text{for each } j \in F \tag{24}$$

$$T_{ik} = 0,1 \quad \text{for each } i = 1,2,...,n \text{ and } k = 0,...,r \tag{25}$$

This model will be called the "best-case" RIP model as it seeks to identify the attack which has the least impact on the system efficiency. This model is based upon the same expected distance values, d_i^k as the "worst-case" RIP model does. The major difference is that we seek to minimize expected weighted distance after an attack on r facilities, where the prob-

ability that any given facility attack is successful is θ. First, the sense of the objective has been reversed to reflect the fact that we want to minimize the expected impact rather than maximize the expected impact by identifying which facilities to attack. The differences between this model and the worst-case RIP model are found in the first three constraints and the objective function. Also, note that now the index k of the variables T_{ik} starts from 0, while in the worst-case RIP k varies between 1 and r. T_{i0} equal to 1 in the new model indicates that the first closest facility to i is not interdicted and, hence, customer i must assign to it. By using this additional set of variables, the inequalities constraint (15) in the worst-case RIP must be turned into the equality constraints (22). This guarantees that, for each i, exactly one of the variables $T_{ik,}$ is equal to 1. Constraints (20) and (21) specify that the variable T_{ik} equal to 1 must be the one associated with the highest value of k such that all of the k closest facilities to i are attacked. Constraint (21) forces the T_{ik} variable to be 1 when all of the k closest facilities to i are attacked and the k+1st closest facility to i is not attacked, thus ensuring that the objective calculates weighted distances associated with the appropriate case for each demand i. "Best-Case" RIP is also an integer programming model and can be solved with general purpose integer-linear programming software.

In this section we have defined two new models that can be used to define the boundaries of expected values of a reliability envelope associated with expected losses of system efficiency when a system is subject to attack and where the success of an attack is specified as a probability θ. In the next section, we give two examples of generating a probabilistic reliability envelope.

11.4 Generating Probabilistic Reliability Envelopes

In the previous section, we developed two new models, each of which characterizes possible losses when the probability of success is less than 1. One model optimizes best-case expected outcomes given a level, r, of interdiction or attacks, and the other identifies worst-case expected losses to system efficiency given a level, r, of interdiction or attacks. In this section we apply the model to the data set of Daskin (1995). The same data set was recently used by Snyder and Daskin (2005) to optimize the location of a set of facilities, while minimizing the cost of expected failure (which differs from worst-case and best-case problems). The set consists of the 150 largest cities in the US according to the 1990 census data. We have used

the largest 100 cities of the 150 city data set to represent demand points. We optimally solved a 10 facility median problem in order to site an existing system. Given an operating system of 10 facilities, we then considered attack or interdiction levels from 1 to 9. We then solved the worst-case RIP and best-case RIP models for two different values of θ. The models were solved by using the CPLEX 9.0 Callable Library on a PC with a Pentium 4, 2.8Ghtz processor and 1GB of RAM. For the problems solved, the solution times were negligible. The average solution time to solve the 9 worst-case RIP problems was 0.1 seconds; the average time for the 9 best-case RIP problems was 0.05 seconds. Although more sophisticated techniques could have been devised to solve the problems, the objective here was to generate example solutions to the model and demonstrate its value in possible infrastructure planning.

Table 11.1 Results of Worst-case RIP and Best-Case RIP models using a probability of an attack being successful of 0.5

		Objective Function Value		Efficiency	
P	***r***	Best-Case	Worst-Case	Best-Case	Worst-Case
10	0	6,913,891,192	6,913,891,192	100%	100%
10	1	7,261,534,325	11,186,485,220	95%	61%
10	2	7,640,648,362	14,160,704,940	90%	48%
10	3	8,069,856,391	16,216,644,966	85%	42%
10	4	8,635,964,514	17,208,987,402	80%	40%
10	5	9,245,683,937	18,475,083,666	74%	37%
10	6	10,201,552,890	19,467,426,102	67%	35%
10	7	11,829,924,971	20,313,229,813	58%	34%
10	8	14,280,633,211	21,509,101,696	48%	32%
10	9	16,892,069,730	22,756,583,295	40%	30%

Table 11.1 gives results when both models, worst-case and best-case RIP, were solved using a probability of θ = .50 (i.e. the probability that any individual attack eliminates or successfully closes a facility). The level or number of facilities subject to attack ranged from r=1 to 9. For each level of attack, the objective function values are given for best-case expected weighted distance and worst-case expected weighted distance. Efficiency for each case is also given as a percentage, where 100% represents the operating level before attack. Notice that as the attack level increases

that the expected outcomes tend to converge. This is due to the fact that as the attack level approaches 10, the number of options between the best-and worst-case tends to decrease. It is also important to note that the greatest marginal impact for worst-case occurs when the attack level is small. Thus, hitting a few key locations in an operating system can have potentially a great impact, especially when the target selections are very intelligent. Figure 11.3 presents the values of expected operation efficiencies (in percent) as a graph, depicting the expected values boundaries of the reliability envelope. Although the best-case efficiencies tend to fall slowly as r increases, it is important to recognize that the greater degree of decision making on the part of an attacker, the closer the outcomes are likely to be to the worst-case expected efficiency curve (i.e. bottom curve of the reliability envelope).

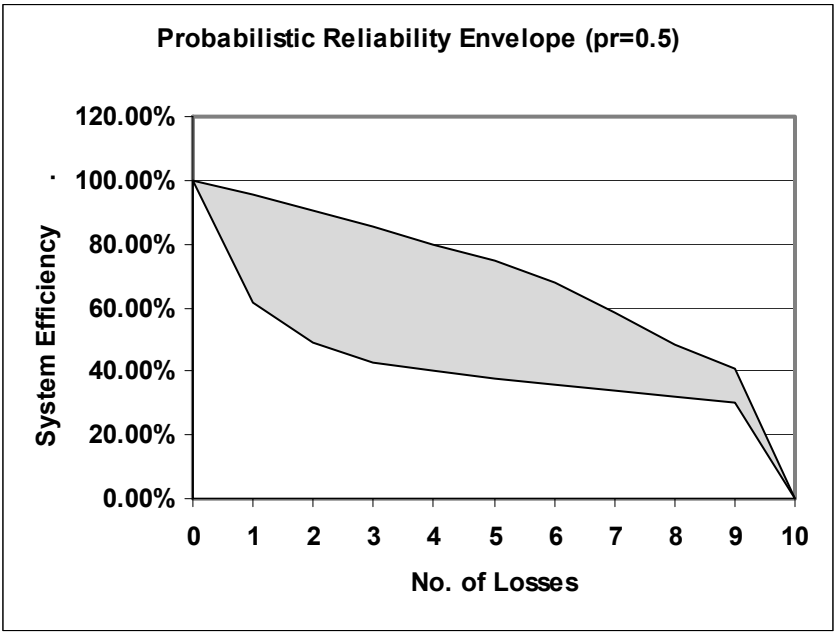


Fig. 11.3 Reliability envelope associated with solutions presented in Table 11.1

We also solved the same set of problems, except that the probability of success for any individual facility attack was set at $\theta = .30$. The results of the 9 different cases of best-case and worse-case RIP models are given in Table 11.2. Expected efficiency levels after attack are much higher than what is given in Table 11.1, since the probability of a successful hit has been decreased. One can see that the impact of having a lower success probability can be substantial. Such information can be very useful in looking at ways to protect a system. Whether the protection is against a terrorist attack or a natural disaster, reducing the probability of success, θ, even by modest amounts could have an impact on system efficiency. For, example, this could be done by placing extra strength in key structural members of a building to protect against an earthquake, or by adding a surveillance system with guards to help protect against an intruder. Either technique may not completely eliminate a loss, by reducing the probability of loss to zero, but such strategies may generate more benefits in terms of improved expected system operating efficiencies (worst-case and best-case) than what it might cost. Thus, the value of this type of analysis could lead to higher levels of safety as well as efficient levels of resource allocation for security measures (whether that involves a possible natural disaster or an attacker). The reliability envelope for the results given in Table 11.2 is presented in Figure 11.4. Note that the expected envelope is thinner than that given in Figure 11.3. As the success probability, θ, decreases, the reliability envelope will tend to narrow.

Table 11.2 Results of Worst-Case RIP and Best-Case RIP models using a probability of an attack being successful of 0.3

		Objective Function Value		Efficiency	
p	***R***	Best-Case	Worst-Case	Best-Case	Worst-Case
10	0	6,913,891,192	6,913,891,192	100%	100%
10	1	7,122,477,072	9,477,447,609	97%	72%
10	2	7,349,945,494	11,087,999,089	94%	62%
10	3	7,607,470,311	11,896,048,109	90%	58%
10	4	7,924,226,980	12,491,453,570	87%	55%
10	5	8,277,928,617	12,909,265,724	83%	53%
10	6	8,840,129,869	13,411,377,650	78%	51%
10	7	9,500,130,641	13,827,865,323	72%	50%
10	8	10,443,063,501	14,329,977,249	66%	48%
10	9	11,857,857,756	14,661,029,043	58%	47%

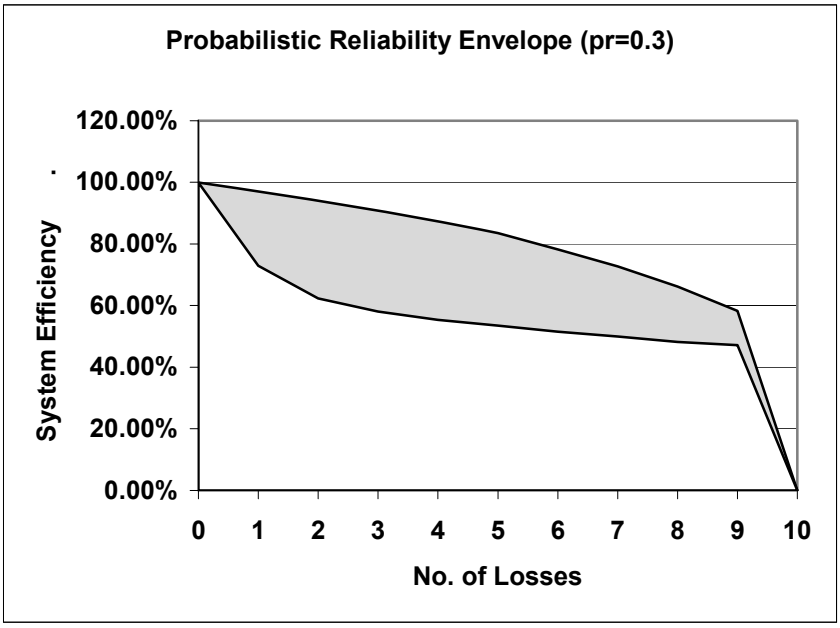


Fig. 11.4. Reliability envelope associate with solution presented in Table 11.2

11.5 Summary and Conclusions

In this chapter, we have described how existing models can be used to generate the reliability envelope for the case where attacks on infrastructure are certain in their outcomes. We know that attacks on infrastructure, whether the result of an interdictor like a terrorist or by events of natural disasters, do not necessarily result in certain losses. In fact, it is reasonable to characterize such losses by a probability that the loss occurs given an attack. Thus, outcomes are less certain and should be expressed in terms of expected outcomes as well as absolute bounds. For the probabilistic case we know that the absolute best-case is that all attacks fail and that efficiency remains at 100%. We also know that the absolute worst-case happens when all attacks are successful. Such a case can be generated by a deterministic model, like RIM given in Section 2. Thus, absolute boundaries of the reliability envelope can be generated by the use of a deterministic loss model. The effective envelope, however, needs to be characterized by

the best-case of expected system efficiency losses and the worst-case expected system efficiency losses.

In this chapter we have also extended the modeling concepts of facility interdiction in order to generate the boundaries of the reliability envelope for the case where the success of attacks on facility infrastructure is probabilistic in nature. We have introduced two new models to solve for the best-case expected interdiction system efficiency and the worst-case interdiction system efficiency, characterized by the number of possible attacks on a system of facilities. Both models are Integer-Linear programming models, which can be solved for modest sized problems using off-the-shelf commercial software. Computation results have been presented for the two models, where two different probabilities of success were used and the resulting envelopes were presented.

The worst-case RIP and best-case RIP models are representative of the types of facility location models that need to be developed to understand the range of impacts that can be inflicted by nature or by an attacker. Research is needed to extend this model framework when facilities are restricted by capacity conditions, where material supplies are intentionally interrupted, and when network elements are possibly compromised. Further, more work is needed in exploring different solution approaches for the both best-case RIP and worst-case RIP.

Acknowledgements

We appreciate the support of the University of Kent in funding portions of this research through its program to encourage collaborative research.

References

Baran, P. 1964. On distributed communication networks. *IEEE Transactions on Communication Systems.* 12,1-9.

Birmbaum, Z.W., J.D. Esary, and S.C. Saunders. 1961. Multi-component systems and structures and their reliability. *Technometrics.* 3, 55-57.

Church, R.L. and J.L, Cohon. 1976. Multiobjective location analysis of regional energy facility siting problems. Report prepared for the U.S. Energy Research and Development Administration (BNL 50567).

Church, R.L., M.P. Scaparra, and R.S. Middleton. 2004. Identifying critical infrastructure: The median and covering facility interdiction problems. *Annals of the AAG.* 94, 491-502.

Church, R.L. 2003. COBRA: a new formulation for the classic *p*-median location problem. *Annals of Operations Research.* 122, 103-120.

Daskin, M.S. 1995. Network and Discrete Location: Models, Algorithms, and Applications. New York: Wiley.

Gerrard, R.A. and R.L. Church. 1996. Closest assignment constraints and location models: Properties and structure. *Location Science.* 4(4), 251-270.

Grubesic, T.H., M.E. O'Kelly, and A.T. Murray. 2003. A geographic perspective on commercial internet survivability. *Telematics and Informatics.* 20, 51-69.

Hakimi, S.L. 1964. Optimum location of switching centers an the absolute centers and medians of a graph. *Operations Research.* 12, 450-459.

Hakimi, S.L. 1965. Optimum distribution of switching centers and some graph related theoretic properties. *Operations Research.* 13, 462-475.

Kim, H. and M. O'Kelly. 2004. Survivability of commercial backbones with peering: a case study of Korean networks. presented at the 51st Annual North American Meetings of the Regional Science Association International, Seattle, WA, Nov. 11-13.

Moore, E.F. and C.E. Shannon. 1956. Reliable circuits using less reliable relays. *J. Franklin Inst.* 262, 191-208.

ReVelle, C. S. and R. Swain. 1970. Central facilities location. *Geographical Analysis.* 2, 30-42.

Rojeski, P.and C.S. ReVelle. 1970. Central facilities location under an investment constraint. *Geographical Analysis* 2, 343-360.

Scaparra, M. P. and R.L. Church. 2006. A bilevel mixed integer program for critical infrastructure protection planning. Working Paper No. 116, Kent Business School, UK.

Snyder, L.V. and M.S. Daskin. 2005. Reliability models for facility location: The expected failure cost case. *Transportation Science. 39*(3), 400-416.

Teitz, M. B. and P. Bart. 1968. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research.* 16, 955-961.

Urban, D. and T. Keitt. 2001. Landscape connectivity: a graph theoretic perspective. *Ecology.* 82, 1205-1218.

12 Bounding Network Interdiction Vulnerability Through Cutset Identification

Timothy C. Matisziw[1], Alan T. Murray[1,2], and Tony H. Grubesic[3]

[1] Center for Urban and Regional Analysis, The Ohio State University, USA; Email: matisziw.1@osu.edu
[2] Department of Geography, The Ohio State University, USA; Email: murray.308@osu.edu
[3] Department of Geography, Indiana University, USA; Email: tgrubesi@indiana.edu

12.1 Introduction

Assessing the vulnerability of network infrastructure to disruptive events is recognized as an important component of network planning and analysis. Motivations for this type of research range from searching for the most effective/ efficient means of disrupting a network (e.g., preventing drug trafficking – see Wood 1993) to assessing possible threats to critical network infrastructures so that adequate protective measures can be devised to limit potential disruption (see Wu 1992). In such analysis, the disruptive activity being examined, whether due to natural disaster, accident, or sabotage, can be generically referred to as network interdiction.

Traditionally, approaches for modeling network interdiction have focused on identifying nodes or linkages most critical to some interpretation of system performance. For instance, increasing the cost associated with routing flow between an origin-destination (O-D) pair is a common goal. Given the objective of increasing transportation costs, the impact of total or partial interdiction of linkages/ nodes can be considered as either: 1) decreasing network capacity, preventing flow or forcing it over more costly alternate paths; or, 2) increasing the cost associated with minimal cost paths. Both aspects of interdiction rely on negatively affecting network connectivity in some way. A classic network analysis approach to impacting connectivity between an O-D pair is through the identification of a cut-

set, or a set of linkages whose removal prevents O-D flow. Provided that interdiction efforts are limited by available resources, it is reasonable to focus on components of the smallest cutset possible (Wood 1993). It has been well established that solution of the maximum-flow model corresponds to a minimum capacity cut; hence, it is no surprise that this relationship has been exploited in the formulation of many interdiction models (Wollmer 1964; McMasters and Mustin 1970; Ghare et al. 1971; Corley and Chang 1974; Ratliff et al. 1975; Cunningham 1985; Phillips 1993; Wood 1993; Burch et al. 2003).

Models based upon a maximum-flow model generally seek to apply limited interdiction resources to minimize the network's capacity to move flow between origins and destinations. To achieve this goal, minimal cutsets can be identified for an O-D pair(s). No other cutset can be contained within a minimal cutset. A minimum capacity cutset then is a cutset of the smallest total weight (however defined). The usefulness of the maximum flow-minimum cut theorem is that the total capacity of a minimum cutset corresponds to the maximum amount of flow capable of moving between an O-D on the network (Ford and Fulkerson 1962; Colbourn 1987; Evans and Minieka 1992). Once minimum capacity cuts are found, linkages in these cuts are likely candidates for attack. The task then becomes determining which component linkages would be interdicted under a budgetary scenario. In this type of model, lower flow capacity remaining in the network indicates a more effective interdiction plan. An algorithmic approach to this problem is presented by Phillips (1993), while Wood (1993) implements this basic idea as an integer program. Though a minimum cutset may indeed be effective for interdiction in certain circumstances, it has been suggested that solution to some problems may require assessment of other minimal cutsets. For instance, if multiple interdiction objectives exist, a minimum capacity cut for each O-D may not necessarily be the most effective option (Boyle 1998; Balcioglu and Wood 2003).

Interdiction of network capacity is indeed an important consideration in assessing a network's vulnerability to interdiction; however, other criteria are also of interest. For instance, how actual origin-destination flow activity may be impacted by interdiction efforts is of obvious concern when addressing network survivability. Discussion on this topic can be found in Wu (1992) and Doyle et al. (2005). More recent analysis of this issue is found in Myung et al. (2004), Matisziw et al. (2006), Grubesic et al. (2006), and Murray et al. (2007). Another fundamental measure of attack vulnerability, therefore, is how network connectivity is impacted by an intentional disruption (see for instance, Holme et al. 2002; Grubesic et al. 2003). The argument is that given an attack on network facilities, higher potential connectivity loss equates to a more vulnerable network infra-

structure. Furthermore, assessment of connectivity underlies the notions of network capacity and flow; hence, interdiction of connectivity is a valid concern when safeguarding network operation. As is clear from the previous discussion, network connectivity is directly related to the concept of minimal cutsets. Obviously, if all elements of a minimal O-D cutset are removed, then connectivity cannot be preserved. From an interdiction standpoint, just as it makes sense to target a minimum capacity cutset, minimum cardinality cutsets are also of interest because they reflect a scenario where limited resources are expended to cause the greatest damage possible. Colbourn (1987) describes one way of deriving minimum cardinality cutsets.

Regardless of the vulnerability measure(s) of concern (e.g., connectivity, capacity, flow), it is vital to understand the outcomes of potential interdiction to better support planning and management of network risk. One way of reducing risk is through the identification of the most disruptive interdiction schemes (those causing maximal damage) to establish an upper bound on vulnerability. If these worst-case scenarios can be identified, then administrators and managers can better plan for protection against threats and system improvement to minimize risks. In fact, many models developed for identifying optimal interdiction plans have their roots in network vulnerability assessment. For instance, the modeling efforts of Wollmer (1964), Corley and Chang (1974), Ratliff et al. (1975); Corley and Sha (1982), Ball et al. (1989), Malik et al. (1989), Church et al. (2004), and Murray-Tuite and Mahmassani (2004) all deal with finding infrastructure components of greatest importance to network operation, or rather the most vital links/ nodes.

However, complete focus on mitigating worst-case damage may not be entirely warranted as many near-optimal interdiction plans may also exist (Grubesic et al. 2006; Matisziw et al. 2006). Evaluation of the range of possible interdiction outcomes is undoubtedly beneficial in this regard, especially if multiple objectives are involved (see Boyle 1998). Hence, aside from an upper (worst-case) performance bound on interdiction severity, establishment of a lower bound (best-case) is also important to guide planning efforts. A higher lower bound may be more indicative of greater interdiction tolerance, as an example. Valid upper and lower bounds can also benefit simulations geared at generating a representative range of potential interdiction outcomes (see Matisziw et al. 2006).

To address the generation of bounds on interdiction of network flows, the flow interdiction model (FIM) has been recently proposed by Murray et al. (2007). The FIM permits assessment of maximally destructive node-based interdiction efforts on network operation. In other words, the FIM can produce an upper bound on the amount of network activity that may be

lost due to a node-based disruption. Interdiction impacts for multiple origins and destinations are easily considered in this modeling framework. As suggested by Murray et al. (2007), the FIM enables either maximization or minimization of flow disruption to be evaluated. This is possible because O-D paths are explicitly tracked.

12.2 Modeling Linkage-based Interdiction

The focus of this chapter is the development of a model capable of producing upper and lower bounds on the loss of connectivity that may result from interdiction efforts aimed at network linkages. In other words, the goal is to identify a cutset of cardinality p that either minimizes or maximizes connectivity of origin and destination pairs. Given an uncapacitated network and the following notation, the p-cutset problem (PCUP) can be formulated:

j = index of linkages, entire set denoted J
k = index of paths, entire set denoted K
o = index of origins, entire set denoted O
d = index of destinations, entire set denoted D
N_{od} = set of paths providing o-d flow
p = number of linkages interdicted
Φ_k = set of linkages along path k

$$X_j = \begin{cases} 1 \text{ if linkage } j \text{ is interdicted} \\ 0 \text{ otherwise} \end{cases}$$

$$Y_k = \begin{cases} 1 \text{ if path } k \text{ remains uneffected by interdiction} \\ 0 \text{ otherwise} \end{cases}$$

$$Z_{od} = \begin{cases} 1 \text{ if no flow possible between } o \text{ - d} \\ 0 \text{ otherwise} \end{cases}$$

p-Cutset Problem (*P*CUP)

$$\textit{Minimize/Maximize} \quad \sum_o \sum_d Z_{od} \tag{1}$$

Subject to:

$$\sum_{k \in N_{od}} Y_k + Z_{od} \geq 1 \qquad \forall o,d \tag{2}$$

$$Z_{od} \leq (1 - Y_k) \qquad \forall\, k \in N_{od} \tag{3}$$

$$Y_k \geq 1 - \sum_{j \in \Phi_k} X_j \qquad \forall k \tag{4}$$

$$Y_k \leq (1 - X_j) \qquad \forall k, j \in \Phi_k \tag{5}$$

$$\sum_j X_j = p \tag{6}$$

$$X_j = \{0,1\} \qquad \forall j \tag{7}$$

$$Y_k = \{0,1\} \qquad \forall k$$

$$Z_{od} = \{0,1\} \qquad \forall o,d$$

Objective (1) is to either minimize or maximize O-D connectivity loss in a network. Constraints (2)-(3) track O-D path availability. Constraints (4)-(5) account for whether a given path is available given the loss of links. The number of linkages to be interdicted is stipulated in Constraint (6). Integer restrictions are specified in Constraints (7).

The PCUP formulation is similar to the FIM detailed in Murray et al. (2007). There are two fundamental differences with the PCUP, however. First, interdiction is considered only for arcs. Second, connectivity is addressed in the PCUP rather than flow.

The PCUP is beneficial in that it permits both minimization and maximization of (1). This is a convenient property since reformulation is not necessary given either goal. One key assumption of the model is that all paths permitting movement or flow between an origin and destination are accounted for. This is necessary for ensuring that a minimal or maximal cutset is identified. That is, here it is assumed that if *any* path connecting an O-D pair exists, then interaction between the two nodes is possible. Otherwise, if *no* path is available, then interaction between the pair cannot occur. Though use of a subset of O-D paths (e.g., *k*-shortest, arc/node disjoint) can reduce problem size, there is no guarantee that an identified cutset is optimal if all O-D paths are not accounted for. The PCUP deals explicitly with total interdiction of linkages (e.g., linkage is either available or is completely disabled) and partial disruption of a linkage is not possible. Worth noting as well is that a special case of the PCUP is the approach proposed in Myung et al. (2004) capable of addressing the maximi-

zation version of the problem. However, in their paper a minimization version of the model is not provided and cannot be obtained via a straightforward extension of their formulation. Furthermore, Myung et al. (2004) propose a heuristically derived bound and only consider a subset of possible O-D paths. The PCUP is an integer program and as such can be solved directly using a commercial optimization package. Here ILOG's CPLEX 6.6 mixed integer optimizer was utilized for solving problem instances. An issue that may arise though is that due to the number of constraints and integer decision variables, achieving optimality may be a computationally demanding task. Murray et al. (2007) discuss some ways in that these issues may be resolved. For example, integer requirements on Y_k and Z_{od} can be relaxed and some constraints can be consolidated (e.g., (5)). Additionally, some constraints could be eliminated from the general model depending on the objective orientation.

12.3 Application of the PCUP

Analysis of *p*-cutsets is conducted on the Abilene Internet2 backbone. The Abilene backbone is a high capacity fiber-optic Internet network connecting member universities within the U.S. (Abilene 2005). The backbone itself consists of 11 routers (nodes) connected by 14 linkages as shown in Figure 12.1.

Here, the PCUP is used to identify those cutsets capable of causing minimal and maximal damage to the network. All nodes in the Abilene network are both origins and destinations of flow and interact with each other. Given this, the network contains 121 interacting O-D pairs. In this network, intra-nodal interaction is present, meaning that flow can move into and out of the same node. Since nodes are not targeted for removal, only 110 O-D pairs (inter-nodal interactions) can potentially be disrupted given link-based interdiction. The O-D paths were obtained by enumerating all simple (loopless) paths for each O-D pair. 896 O-D exist, requiring approximately 2 seconds of computational time. Both the maximization and minimization cases are examined here for a range of interdiction scenarios.

Table 12.1 and Figure 12.2 illustrate results maximizing O-D connectivity loss. Since every node in the Abilene backbone is directly connected to at least two other nodes (a 2-degree node), the interdiction of a single linkage can not disconnect any O-D pair. However, when two linkages are rendered inoperative, more than half (60) of the O-D pairs lose connectivity. For example, the PCUP identifies the Kansas City-Indianapolis and

Houston-Atlanta linkages in the 2-cutset (p=2) causing the greatest impact. This cutset essentially partitions the network such that the number of nodes

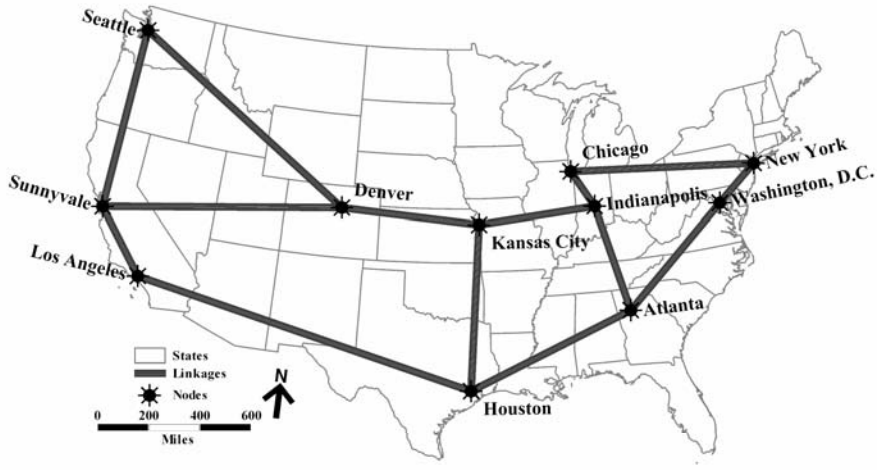


Fig. 12.1 Abilene Internet2 network backbone

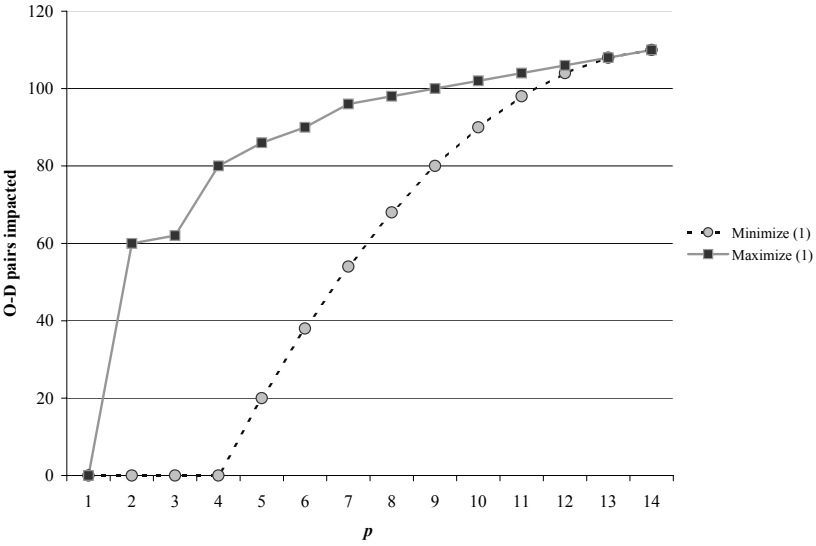


Fig. 12.2 Connectivity impact for minimal and maximal p-cutsets

in each half is as balanced as possible, thereby maximizing disruption. For p=3, the PCUP identifies the Sunnyvale-Los Angeles, Denver-Kansas City, and Los Angeles-Houston linkages as a 3-cutset causing maximum connectivity loss. Los Angeles is consequently disconnected from the net-

work in this instance (see Table 12.1). Figure 12.3 shows the maximum impact of a 7 linkage failure/attack on the backbone. This particular interdiction plan fragments the network into 5 components, disconnecting all but 14 O-D pairs. Given a linkage-based interdiction plan, the maximum number of O-D pairs that can be interdicted is 110 since intra-nodal interaction cannot be impacted by a linkage-based attack. Thus, some 87% of O-D flow interactions are impacted in this case.

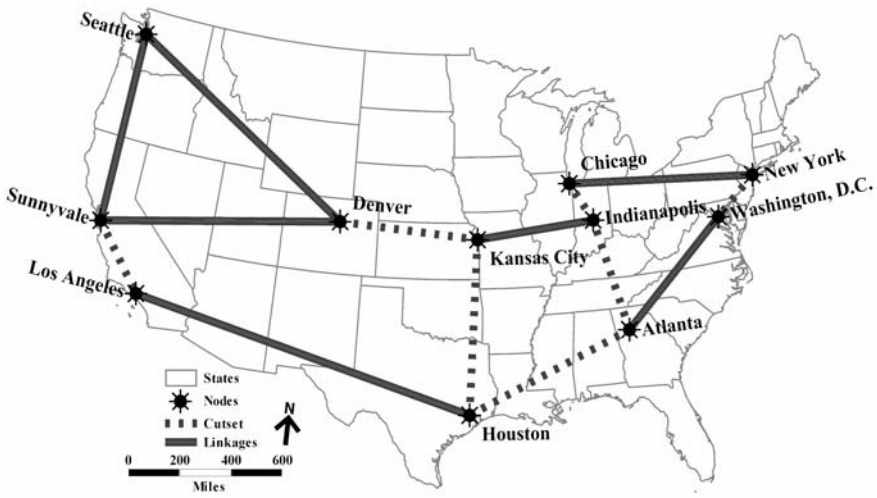


Fig. 12.3 Maximum 7-cutset (p=7)

Minimization of connectivity loss produces very different outcomes. Analysis for the minimization case of the PCUP is presented in Table 12.2 and Figure 12.2. The results presented in Table 12.2 illustrate the best-case situations for system performance given the occurrence of each interdiction scenario. In these cases, the model tries to preserve connectivity between the O-D pairs to the greatest extent possible in a rather intuitive manner. Given a set of nodes V in a network, it is well-known that minimum network connection occurs when $|J| = (|V|-1)$. In the case of the Abilene backbone there are 11 nodes, so a minimum of 10 linkages are needed to maintain connectivity. Since the backbone is composed of 14 linkages, up to four can be removed without causing connectivity loss. This is exactly the result found using the PCUP. After enough linkages are eliminated to reduce the network to a spanning tree, then reduction of each additional linkage disconnects exactly one node from the network, retaining a minimally connected network between the remaining nodes. That is, for values of $p > |J|-(|V|-1)$ in an undirected network where all nodes interact with each other, the maximum connectivity remaining in the network

Table 12.1 Maximizing connectivity loss

p	OB	IT	BR	TM*	Linkages Interdicted
1	0	1301	10	9.890	CH-NY
2	60	348	0	1.391	IN-KC;AT-HO
3	62	1649	28	11.156	DE-KC;SV-LA;LA-HO
4	80	285	0	1.046	DE-KC;CH-IN;LA-HO;WA-AT
5	86	181	2	1.938	DE-KC;CH-IN;SV-LA;LA-HO;WA-AT
6	90	240	4	2.406	DE-KC;CH-IN;SV-LA;LA-HO;WA-AT;CH-NY
7	96	163	8	2.688	DE-KC;CH-IN;SV-LA;KC-HO;AT-HO;IN-AT;NY-WA
8	98	240	15	2.438	DE-KC;CH-IN;SV-LA;KC-HO;AT-HO;IN-AT;NY-WA;CH-NY
9	100	48	0	0.453	DE-KC;CH-IN;SE-SV;DE-SV;LA-HO;KC-HO;IN-AT;WA-AT;NY-WA
10	102	49	0	0.469	DE-KC;CH-IN;SE-SV;DE-SV;LA-HO;KC-HO;AT-HO;IN-AT;WA-AT;NY-WA
11	104	42	0	0.469	DE-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT;WA-AT;NY-WA
12	106	42	0	0.562	DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT;WA-AT;NY-WA
13	108	41	0	0.453	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT;WA-AT;NY-WA
14	110	0	0	0.234	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT;WA-AT;NY-WA;CH-NY

OB Objective, *IT* Iterations, *BR* Branches, *TM* Time, *AT* Atlanta, *CH* Chicago, *DE* Denver, *HO* Houston, *IN* Indianapolis, *KC* Kansas City, *LA* Los Angeles, *NY* New York, *SE* Seattle, *SV* Sunnyvale, *WA* Washington, D.C.
*Solution times in seconds for Pentium III parallel processor with 1.0 GB RAM

given the removal of p linkages can be determined as follows: 1) compute the number of nodes that become separated from the network $V_S = p-(|J|-(|V|-1))$, 2) compute the number of nodes retained in the network $V_R=(|V|-V_S)$, 3) since the network design preserving connectivity is a tree, it is known that $V_R(V_R-1)$ node pairs will remain connected, and 4) the difference between original network connectivity and that remaining after p linkages are removed gives the connectivity lost (PCUP's objective). Figure 12.4 illustrates the minimum connectivity loss resulting from a 7-cutset interdiction, in contrast to the maximum scenario shown in Figure 12.3. Note that through a minimum network connection, interaction between 8 nodes (56 O-D pairs) can be preserved. While identifying a spanning tree may be an alternative and attractive way of solving this minimization problem, it is unclear whether such a technique will always result in an optimal solution if: 1) all nodes do not interact with each other, and/or 2) actual O-D flow activity is incorporated within the model.

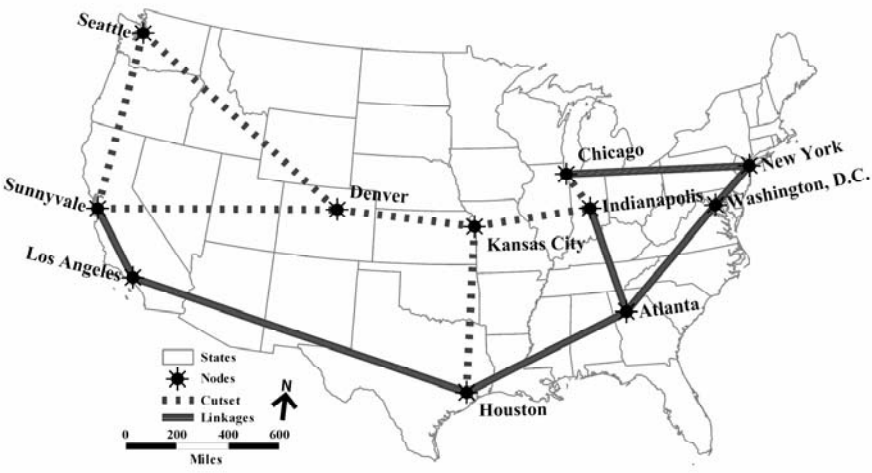


Fig. 12.4 Minimum 7-cutset (p=7)

In the existing literature, it is commonly assumed that individual characteristics of network nodes or linkages (e.g., degree) can be used as proxies to infer the importance of a facility to network operation (Holme et al. 2002, for example). However, the result of an interdiction is in fact strongly dependant upon the spatial structure of the network, as can be seen in the model results. As an example, Table 12.2 shows that a 2-cutset resulting in minimum O-D disruption involves the Denver-Kansas City and Atlanta-Houston linkages that are both rooted by different nodes of degree 3. In this case, no O-D connectivity is lost. On the other hand, the

Table 12.2 Minimizing connectivity loss

p	OB	IT	BR	TM*	Linkages Interdicted
1	0	915	0	4.375	DE-KC
2	0	667	1	23.344	DE-KC;AT-HO
3	0	433	1	22.141	LA-HO;AT-HO;IN-AT
4	0	356	1	12.547	SE-DE;DE-SV;IN-AT;WA-AT
5	20	96118	1246	151.953	DE-KC;IN-KC;CH-IN;DE-SV;CH-NY
6	38	110907	1891	195.063	SE-DE;DE-KC;IN-KC;CH-IN;DE-SV;IN-AT
7	54	88637	1649	179.047	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;KC-HO
8	68	60221	1346	140.125	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO
9	80	30563	602	85.609	IN-KC;CH-IN;DE-SV;KC-HO;AT-HO;IN-AT;WA-AT;NY-WA;CH-NY
10	90	17946	325	66.250	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO
11	98	10537	157	53.125	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT
12	104	5858	91	58.235	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT;WA-AT
13	108	2694	22	41.219	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT;NY-WA;CH-NY
14	110	0	0	0.235	SE-DE;DE-KC;IN-KC;CH-IN;SE-SV;DE-SV;SV-LA;LA-HO;KC-HO;AT-HO;IN-AT;WA-AT;NY-WA;CH-NY

OB Objective, *IT* Iterations, *BR* Branches, *TM* Time, *AT* Atlanta, *CH* Chicago, *DE* Denver, *HO* Houston, *IN* Indianapolis, *KC* Kansas City, *LA* Los Angeles, *NY* New York, *SE* Seattle, *SV* Sunnyvale, *WA* Washington, D.C.
*Solution times in seconds for Pentium III parallel processor with 1.0 GB RAM

2-cutset identified in Table 12.3 (discussed above) also involves different nodes of degree 3. However, their selection forms a spatial partition, maximizing O-D connectivity loss. While both of these 2-cutsets have similar physical characteristics, their impact on network functionality (if interdicted) is not at all similar. In fact, these two cutsets form upper and lower bounds on possible connectivity loss due to 2-link interdiction and serve to illustrate the range of interdiction outcomes possible. Measures based on proxies for connectivity are not likely to be good approximations for these bounds and could obscure the true extent of network vulnerability. Nonetheless, many other feasible interdiction outcomes undoubtedly occur for each interdiction scenario (e.g., p=2) and may include near-optimal solutions or alternate-optima. Hence, from the perspective of managing network vulnerabilities, there is still a clear benefit in characterizing the range of possible outcomes between the upper and lower bounds through simulation as discussed in Matisziw et al. (2006).

12.4 Discussion and Conclusion

This chapter has focused on identifying minimal or maximal p-cutsets for a system of origins and destinations. The goal is to obtain the set of p cardinality cuts, or linkages, capable of maximizing or minimizing network connectivity loss. This distinction is important in that other models have focused primarily on the interdiction of capacity, not connectivity or other measures of network vulnerability. Furthermore, models that have approached connectivity have typically done so using proxies for connectivity (e.g., nodal degree, betweeness, etc.) and have not modeled it exactly as is done here. The motivation for this problem follows directly from the need to assess a network's vulnerability to interdiction. Effective planning and management of network risks must consider the range of interdiction scenarios possible if appropriate mitigation measures are to be devised.

Recent events emphasize the importance of such analysis. For instance, single-link failures are a common occurrence in the operation of many networks and hence a common consideration in network design (Wu et al. 1988). Although many networks are resilient to single-link attack/ failure, additional, simultaneous disruptions can be very problematic, leading to wide-spread service outages. A recent example of this type of service outage is that caused by a 2-cut in the Sprint Nextel fiber optic network in the southwestern U.S. (C|net 2006; CNN 2006).

In order to identify the upper and lower bounds on post-interdiction network connectivity, the p-cutset problem (PCUP) was proposed. The

PCUP is an extension of the flow interdiction model of Murray et al. (2007) aimed at explicitly accounting for linkage-based interdiction. Unlike other interdiction models, the PCUP's structure permits both maximization and minimization of network connectivity loss. This is accomplished by enumeration of O-D paths, permitting identification of cutsets of a stipulated cardinality that disconnect or preserve the greatest number of O-D relationships. Through application to a real world network, the PCUP is shown to be effective for identifying bounds on potential interdiction scenarios.

Acknowledgements

Project funding for Matisziw and Murray is provided through the Center for Urban and Regional Analysis at The Ohio State University.

References

Abilene. 2005. http://abilene.internet2.edu.

Balcioglu, A. and R.K. Wood. 2003. Enumerating near-min s-t cuts. In *Network Interdiction and Stochastic Integer Programming*. Edited by D.L. Woodruff. Boston: Kluwer Academic Publishers, 51-69.

Ball, M.O., B.L. Golden, and R.V. Vohra. 1989. Finding the most vital arcs in a network. *Operations Research Letters*. 8(2), 73-76.

Burch, C., R. Carr, S. Krumke, M. Marathe, C. Phillips, and E. Sundberg. 2003. A decomposition-based approximation for network inhibition. In *Network Interdiction and Stochastic Integer Programming*. Edited by D.L. Woodruff. Boston: Kluwer Academic Publishers, 51-69.

Boyle, M.R. 1998. *Partial-Enumeration for Planar Network Interdiction Problems*. M.S. Thesis: Naval Postgraduate School.

C|net News.com. 2006. Sprint Nextel suffers service outage. http://www.news.com. Monday, Jan. 9.

Church, R.L., M.P. Scaparra, and R.S. Middleton. 2004. Identifying critical infrastructure: the median and covering facility interdiction problems. *Annals of the Association of American Geographers*. 94(3), 491-502.

Colbourn, C.J. 1987. *The Combinatorics of Network Reliability*. New York: Oxford.

Corley, H.W. and H. Chang. 1974. Finding the *n* most vital nodes in a flow network. *Management Science*. 21, 362-364.

Corley, H.W. and D.Y. Sha. 1982. Most vital links and nodes in weighted networks. *Operations Research Letters*. 1(4), 157-160.

CNN.com. 2006. Cut cable quiets Sprint service in West. http://www.cnn.com. Jan. 9.

Cunningham, W.H. 1985. Optimal attack and reinforcement of a network. *Journal of the Association for Computing Machinery*. 32(3), 549-561.

Doyle, J.C., D.L. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger. 2005. The "Robust Yet Fragile" Nature of the Internet. *Proceedings of the National Academy of Sciences of the United States of America*. 102(4), 14497-14502.

Evans, J. and E. Minieka. 1992. *Optimization Algorithms for Networks and Graphs*. 2nd Edition. New York: Marcel Dekker, Inc.

Ford, L.R. and D.R. Fulkerson. 1962. *Flows in Networks*. Princeton Press.

Ghare, P.M., D.C. Montgomery, and W.C. Turner. 1971. Optimal interdiction policy for a flow network. *Naval Research Logistics Quarterly*. 18, 37-45.

Grubesic, T.H., M.E. O'Kelly, and A.T. Murray. 2003. A geographic perspective on commercial internet survivability. *Telematics and Informatics*. 20, 51-69.

Grubesic, T.H., T.C. Matisziw, and A.T. Murray. 2006. Targeted attacks and survivability in critical network infrastructure. *Submitted for publication*.

Holme, P., B.J. Kim, C.N. Yoon, and S.K. Han. 2002. Attack vulnerability of complex networks. *Physical Review E*. 65, 056109.

Malik, K., A.K. Mittal, and S.K. Gupta. 1989. The most vital arcs in the shortest path problem. *Operations Research Letters*. 8, 223-227.

Matisziw, T.C., A.T. Murray, and T.H. Grubesic. 2006. Exploring the vulnerability of network infrastructure to interdiction. *Submitted for review*.

McMasters, A.W. and T.M. Mustin. 1970. Optimal interdiction of a supply network. *Naval Research Logistics Quarterly*. 17, 261-268.

Murray, A.T., T.C. Matisziw, and T.H. Grubesic. 2007. Critical network infrastructure analysis: interdiction and system flow. *Journal of Geographical Systems*, 39.

Murray-Tuite, P.M. and H.S. Mahmassani. 2004. Methodology for determining vulnerable links in a transportation network. *Transportation Research Record*. 1882, 88-96.

Myung, Y-S. and H. Kim. 2004. A cutting plane algorithm for computing *k*-edge survivability of a network. *European Journal of Operational Research*. 156, 579-589.

Phillips, C.A. 1993. The network inhibition problem. *Proceedings of the Annual Association for Computing Machinery STOC*. California, May.

Ratliff, H.D., G.T. Sicilia, and S.H. Lubore. 1975. Finding the *n* most vital links in flow networks. *Management Science*. 21(5), 531-539.

Wollmer, R. 1964. Removing arcs from a network. *Operations Research*. 12, 934-40.

Wood, R. K. 1993. Deterministic network interdiction. *Mathematical Computer Modelling*. 17(2), 1-18.

Wu, T-H. 1992. *Fiber Network Service Survivability*. Boston: Artech House.

Wu, T-H., D.J. Kolar, and R.H. Cardwell. 1988. Survivable network architectures for broad-band fiber optic networks: model and performance comparison. *Journal of Lightwave Technology*. 6(11), 1698-1709.

13 Models for Reliable Supply Chain Network Design

Lawrence V. Snyder[1], Mark S. Daskin[2]

[1]Department of Industrial and Systems Engineering, Lehigh University, USA; Email: larry.snyder@lehigh.edu
[2]Department of Industrial Engineering and Management Sciences, Northwestern University, USA; Email: m-daskin@northwestern.edu

13.1 Introduction

Recent examples of disruptions in the news suggest a strong geographical dimension to supply chain disruptions, and to their effects. For example:

- The west-coast port lockout in 2002 strangled U.S. retailers' supply lines while east-coast ports were essentially unaffected (Greenhouse 2002)
- The foot-and-mouth disease scare in the U.K. in 2001 caused the U.S. to ban imports of British meat (Marquis and McNeil 2001).
- The suspension of the license of the Chiron plant in Liverpool, England reduced the U.S. supply of the influenza vaccine by nearly 50% during the 2004/5 flu season (Pollack 2004).
- In the U.S. Gulf Coast region in 2005, Hurricane Katrina idled facilities situated at all levels of the supply chain, including production (e.g., coffee; Barrionuevo and Deutsch 2005), processing (oil refining; Mouawad 2005), warehousing (lumber storage; Reuters 2005), transit (banana imports; Barrionuevo and Deutsch 2005), and retail (groceries and home-repair; Fox 2005, Leonard 2005). These facilities were located in or near New Orleans but were integral parts of global supply chains.

These examples highlight the need for supply chain design models that account for the spatial nature of both supply chains and their operation.

In this chapter, we present several models for reliable facility location in a supply chain that is vulnerable to disruptions. Since facility location decisions are costly to implement and difficult to reverse, these strategic decisions permit very little recourse once a disruption occurs, other than reassignment of customers to non-disrupted facilities. Our goal, therefore, is to choose facility locations proactively so that the system performs well even if disruptions occur.[1]

Consider the following example. Fig. 13.1 depicts the optimal solution to the uncapacitated fixed-charge location problem (UFLP) for a 49-node data set consisting of the capitals of the 48 continental U.S. states and Washington, DC. All nodes serve as both potential facility location sites and demand points, with demands proportional to state populations. This data set is modified from Daskin (1995). The optimal UFLP solution entails a fixed cost of $386,900 per year to operate the five opened facilities and a transportation cost of $470,228 per year.

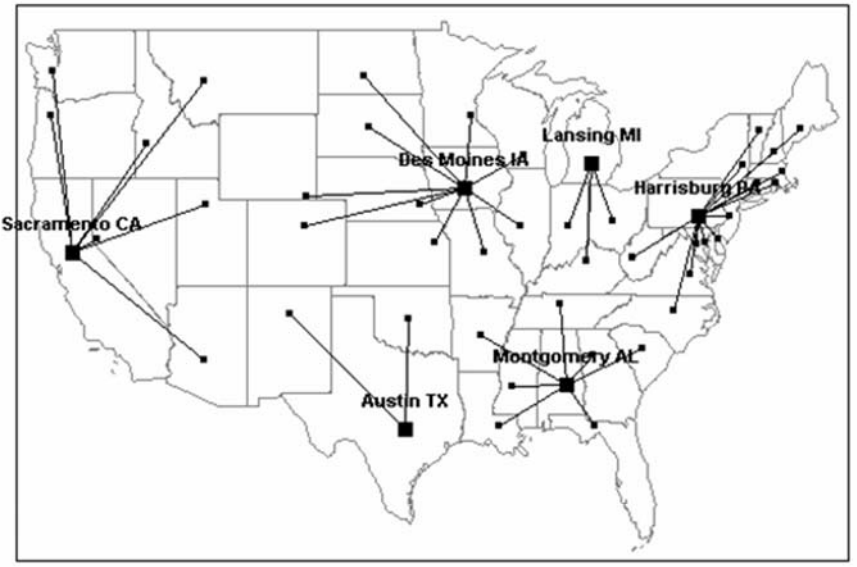


Fig. 13.1. UFLP solution for 49-node dataset

Now suppose that the facility in Sacramento, California becomes unavailable—say, because of a strike or extended power outage. In this case,

[1] In this chapter, we use the terms "failure" and "disruption" interchangeably.

the west-coast customers served by that facility must instead be served by facilities in Des Moines, Iowa and Austin, Texas (Fig. 13.2), resulting in a transportation cost of $1,019,065, an increase of 117% from the baseline solution.

Table 13.1 lists the "failure costs" (the transportation costs that result after the failure of a facility) for each of the five facilities in the optimal solution, as well as their assigned demands and the transportation cost when no facilities fail. Note that Sacramento serves only 19% of the total demand but generates the largest failure cost because its customers are geographically disparate and the next-closest facility is quite distant. The Harrisburg facility serves customers that are tightly clustered, and good "backup" facilities are fairly close by, but its failure cost is still quite large (a 52% increase in transportation cost) because of the volume of demand that it serves. In contrast, Montgomery serves nearly as much demand as Sacramento, but because it is centrally located, close to backup facilities, its failure cost is smaller than that of Sacramento or Harrisburg. Therefore, the reliability of a facility depends on both the demand served by the facility and the distance of those demands from other facilities.

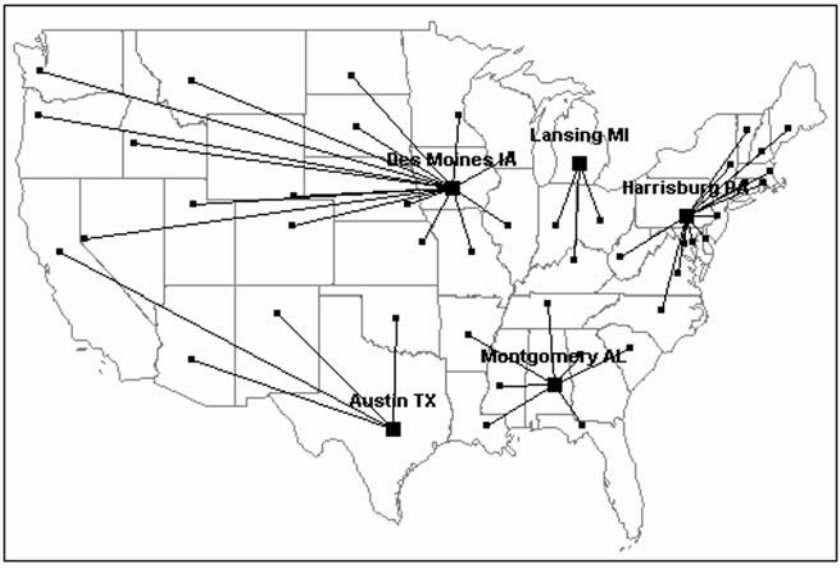


Fig. 13.2. UFLP solution for 49-node dataset, after failure of facility in Sacramento

Table 13.1. Failure costs and assigned demands for UFLP solution

Location	% Demand Served	Failure Cost	% Increase
Sacramento, CA	19	1,019,065	117
Harrisburg, PA	29	713,482	52
Montgomery, AL	17	634,473	35
Austin, TX	9	593,904	26
Des Moines, IA	16	546,599	16
Lansing, MI	12	537,347	14
Transportation cost w/o failures		470,228	0

A more reliable solution locates facilities in the capitals of Alabama, California, Iowa, New York, Ohio, Oregon, Pennsylvania, and Texas. The maximum failure cost occurs when the Austin, TX fails, but this cost is only $476,374, a mere 35 percent increase over the transportation cost of $352,698 when all 8 facilities are working. On the other hand, this solution also requires two additional facilities and is suboptimal for the UFLP. This solution is 7 percent more expensive according to the classical measure of cost (the UFLP cost) but is less expensive when failures are accounted for.

We argue that this latter measure (accounting for failures) is a more accurate measure of cost and that the second solution may be preferable to the first because of its superiority in this measure. Indeed, one of the key aims in this chapter is to demonstrate that large improvements in reliability can often be attained with only small increases in the classical cost.

Although we believe strongly that the "correct" objective functions in facility location problems should account for failures, we also believe strongly that it is important to examine the tradeoff between this objective and the classical ones—that is, the tradeoff between the cost if no disruptions occur and the cost if disruptions do occur. This tradeoff allows us to determine how significant a cost increase is required to add reliability to a system. For example, normal operating cost (sum of the fixed plus transportation costs) had to be twice as large as the optimal UFLP cost to attain a reasonable level of reliability, the additional cost may be unwarranted (unless facility failures are very likely). If, on the other hand, the tradeoff curve is "steep," then firms do not need huge investments in redundant infrastructure to improve the system's reliability. We believe that developing such tradeoff curves is an important step in convincing firms to change their optimization objectives to include disruptions.

Indeed, we generally find that the tradeoff curve is steep in this way. One explanation for this fortuitous finding is that, like many combinatorial optimization problems, facility location problems tend to have many near-optimal solutions. Some of these solutions may, by chance, have desirable

properties like reliability. If we can find these solutions, we may find that the their attractive properties outweigh their slight suboptimality.

Of course, there are a number of possible ways to formulate objectives that consider disruptions. For example, one might try to minimize the expected failure cost (by weighting the failure costs in Table 13.1 by the probability of each facility's disruption), minimize the maximum failure cost (among all rows in Table 13.1), or find a solution whose cost stays within a given threshold with some probability.

In this chapter, we consider optimization models for the design of reliable facility location systems under a variety of risk measures and operating strategies, including those discussed in the previous paragraph and others. Our focus is on the formulation of these models and the insights that can be gained from comparing solutions obtained from different objectives. We briefly discuss algorithmic techniques for solving some of these models, but generally we refer to other sources for such discussions.

The remainder of this chapter is organized as follows. We present a brief literature review in Sect. 2. In Sect. 3, we introduce a base model that will be used as a foundation for the other models to follow. We discuss two ways to formulate this model, as well as a capacitated extension. In Sect. 4, we formulate several models using a range of risk measures. We summarize our findings and discuss opportunities for future research on Sect. 5.

13.2 Literature Review

In this section, we present a brief overview of the literature on reliable supply chain network design problems. A more formal review of this body of literature is presented by Snyder et al. (2006). We refer the reader to the textbooks by Daskin (1995), Drezner, (1995), or Drezner and Hamacher (2002) for an introduction to facility location. Owen and Daskin (1998), Daskin, Snyder, and Berger (2005), and Snyder (2006) all provide reviews of stochastic location models (generally considering uncertainty in demand, rather than disruptions to facilities). See Birge and Louveaux (1997) or Higle (2005) for an introduction to general stochastic programming techniques.

Snyder and Daskin (2005) introduce several models, based on classical facility location problems, in which facilities may fail with a given probability. They minimize a weighted sum of two objectives, one of which is a classical objective (ignoring disruptions) and the other of which is the expected cost after accounting for disruptions. Customers are assigned to

several facilities, one of which is the "primary" facility that serves it under normal circumstances, one of which serves it if the primary facility fails, and so on. One of their models is discussed below in Sect. 13.3.2. Snyder and Ülker (2005) present a capacitated version of their model (Sect. 13.3.1) and Jeon, Snyder, and Shen (2006) present a version that incorporates inventory costs into the location decision.

Berman, Krass, and Menezes (2005a) consider structural properties of a model that is less computationally tractable than Snyder and Daskin's but more general. A subsequent paper (Berman, Krass, and Menezes 2005b) assumes that customers do not know in advance which facilities are operational and must travel from facility to facility in search of a working site.

Church and Scaparra (2005) and Scaparra and Church (2005, 2006) consider the fortification, rather than design, of facilities—that is, the network is assumed to exist and the firm has resources to prevent disruptions at some of them, thus partially fortifying the network. Their model finds the best facilities to fortify assuming that an interdictor will attempt to cause worst-case losses for the firm by disrupting a fixed number of the un-fortified facilities. Similarly, Daskin et al. (2005) allow the firm to choose whether each facility opened is reliable or unreliable; reliable facilities come at a higher cost. (See Sect. 13.4.2 below).

Reliable facility location models are related to network reliability theory (Coburn 1987, Shier 1991, Shooman 2002), which attempts to calculate or maximize the probability that a network remains connected after random link failures. It is also related to the literature on facility location with congestion, in which facilities are sometimes unavailable due to excess demand (rather than to facility disruptions). (See Berman and LeBlanc (1984), Berman et al. (1985), Daskin (1982, 1983), Larson (1974), ReVelle and Hogan (1989).)

13.3 Base Model

In this section, we present a base model that will be used as a foundation for most of the models to come. We formulate this base model in two ways. The first method uses scenarios to represent uncertain events and resembles the formulation of other stochastic facility location problems. This formulation is quite flexible and can be used to model the variations discussed throughout this chapter. However, the number of scenarios may be exponentially large: If there are N facilities and each can fail independently, there are 2^N failure scenarios. This type of formulation was used previously for a capacitated facility location problems with disruptions

(Snyder and Ülker 2005). We present an uncapacitated version first, and then the capacitated version.

The second method captures the uncertain events implicitly, without explicit enumeration of all failure scenarios, and can be solved more efficiently than the scenario-based formulation. Unfortunately, it requires a restrictive assumption (that all facilities have the same probability of disruption) and cannot be extended with the same flexibility as the scenario-based formulation. This formulation was first introduced by Snyder and Daskin (2005a).

All of our models are based on the uncapacitated fixed-charge location problem (UFLP; Balinski 1965, Daskin 1995). We are given a set I of customer locations, each of which has an annual demand h_i for a single product. In addition, we have a set J of potential facility sites, each with an annual fixed operating cost f_j. If we choose to open facility j, then f_j is incurred at all times, regardless of whether the facility is operational. The cost to transport one unit of demand from facility j to customer i is denoted d_{ij}.

In the classical UFLP, there are two sets of decision variables, location variables and assignment variables. The location variables are denoted by X_j, which equals 1 if we open a facility at site j. The formulation of the assignment variables is different for different models below; we defer further discussion until we formulate those models.

Associated with each customer is a per-unit penalty cost θ_j that represents the cost of not serving the customer. This cost is incurred if all open facilities have failed, or if the facilities close to i (with respect to the transportation cost d_{ij}) have failed so that it is cheaper to pay the penalty than to serve the customer. θ_j may represent a lost-sales cost, or the cost to pay a competitor to serve the customer temporarily. Rather than modeling this cost explicitly, we add a dummy "emergency facility," denoted u, to the set J. Facility u is always open, has no fixed cost, and has a transportation cost of θ_j to customer i—that is, $X_u = 1$, $f_u = 0$, and $d_{iu} = \theta_j$ for all i. Moreover, facility u can never fail. Henceforth, we assume that the facility set J has been augmented in this way, and we ignore the penalty cost θ_j.

13.3.1 Scenario-Based Formulation

Model

Let S be a set of scenarios, each of which specifies the failure state of all facilities in J. In particular, let A_s be the set of facilities that fails in sce-

nario s. For convenience, we also define $a_{js} = 1$ if facility j fails in scenario s and 0 otherwise. Scenario s occurs with probability q_s. These scenarios may have been identified *a priori* by managers as likely possibilities that are worth planning against. Alternately, they may represent *all* possible combinations of facility failures. For example, if each facility j fails with probability p_j and failures are independent, then scenario s occurs with probability

$$q_s = \prod_{j \in A_s} p_j \prod_{j \in J \setminus A_s} (1 - p_j). \tag{1}$$

We can modify these probabilities accordingly if failures are dependent. (Failures may be dependent because of geographic proximity, supplier commonality, etc.) To model the emergency facility, we require $a_{us} = 0$ for all s, or, equivalently, $q_s = 0$ if $a_{us} = 1$.

The scenario probability q_s is interpreted as the long-run fraction of time that the precise set of facilities A_s is disrupted. Put another way, the fraction of time in which facility j is disrupted is given by $p_s = \sum_{s \in S: j \in A_s} q_s$. In some cases, the q_s may be estimated from historical data, while in others it must be estimated subjectively. Our models are most easily interpreted as infinite-horizon models in which the facilities in A_s are disrupted for q_s fraction of the time. However, if the modeler has in mind a particular finite time horizon T, then q_s may be used to capture probabilistic information about the timing of the disruptions.

For example, suppose scenario s represents the situation in which exactly one facility, j, fails. Further, suppose that facility j will fail with probability 0.1, and if it does, it will fail in all periods from 1 through 5 with probability 0.3 and in all periods from 3 through T with probability 0.7. (Note that this means that if j fails at all, it will surely be non-operational during periods 3 through 5.) Then q_s is given by

$$q_s = \frac{0.9 \times 0 + 0.1 \times [0.3 \times 5 + 0.7 \times (T - 2)]}{T}. \tag{2}$$

For simplicity, we assume that scenarios specify only facility failures. However, it is simple to extend this formulation so that demands and transportation costs are also scenario dependent.

In each scenario, we need to assign customers to facilities. The decision variable for these doing so is given by Y_{ijs}, which equals the fraction of customer i's demand that is assigned to facility j in scenario s. As in the classical UFLP, single sourcing is optimal; that is, there exists an optimal solution for which $Y_{ijs} \in \{0,1\}$ for all i, j, and s.

We formulate our base model with the objective of minimizing the expected cost, though in future sections we will consider alternate risk measures. The scenario-based formulation of the reliability fixed-charge location problem (RFLP1) is formulated as follows:

$$\text{(RFLP1)} \quad \text{minimize} \sum_{j\in J} f_j X_j + \sum_{s\in S}\sum_{i\in I}\sum_{j\in J} q_s h_i d_{ij} Y_{ijs} \tag{3}$$

$$\text{subject to} \sum_{j\in J} Y_{ijs} = 1 \qquad \forall i \in I, s \in S \tag{4}$$

$$Y_{ijs} \le (1 - a_{js}) X_j \qquad \forall i \in I, j \in J, s \in S \tag{5}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{6}$$

$$Y_{ijs} \ge 0 \qquad \forall i \in I, j \in J, s \in S \tag{7}$$

The objective function (3) minimizes the fixed cost plus the expected transportation cost across all scenarios. Constraints (4) require each customer to be assigned to some facility in every scenario. Constraints (5) prohibit a customer from being assigned to a facility that has not been opened, or to a facility that has failed in a given scenario. Constraints (6) require the location variables to be binary, and constraints (7) require the assignment variables to be non-negative (though, as stated above, an optimal solution always exists in which they are binary). Note that, although we do not explicitly require $X_u = 1$, any optimal solution will open the emergency facility if it is needed for some scenario since it has no fixed cost.

Note that, if there is a single scenario, and no facilities fail in this scenario, this model reduces to the classical UFLP. Since the UFLP is NP-hard (Garey and Johnson 1979), so is the RFLP.

(RFLP1) can be solved using standard IP solvers like CPLEX. However, if the scenarios represent all possible combinations of failures, then S is exponentially large. In this case, sampling techniques such as sample average approximation (SAA; ; Kleywegt, Shapiro and Homem-de-Mello 2001; Linderoth, Shapiro, and Wright 2002) may be used to solve the problem with a reduced set of scenarios and obtain statistical bounds on the quality of the solutions.

Capacitated Model

The formulation above assumes that facilities have infinite capacity or that they can serve any number of demands. In many cases, this might not be true. We can define k_{js} to be the capacity of a facility at candidate site j in scenario s. This notation and the following formulation, allow a facility to incur impaired capacity in a scenario without completely failing. We let the capacity of the dummy facility u be $k_{us} = \infty$ for all scenarios s, indicating that this facility can accommodate all demands if necessary in each scenario. With this notation, we replace constraint (5) by its more traditional version

$$Y_{ijs} \le X_j \qquad \forall i \in I, j \in J, s \in S \tag{8}$$

In addition, we add the following capacity constraint, where the demand placed on a facility's capacity is measured in terms of the demand units h_i

$$\sum_{i \in I} h_i Y_{ijs} \le k_{js} X_j \qquad \forall j \in J, s \in S \tag{9}$$

This formulation, denoted CRFLP, was first suggested by Snyder and Ülker (2005).

Two observations are worth making about the CRFLP. First, constraints (8) are implied by (9) and are therefore not technically needed. However, in most cases, the addition of (8) will strengthen any relaxation of the model. Hence, we suggest including constraints (8) explicitly in any model or algorithm. Second, constraints (9) allow demands at a node to be split between multiple facilities since the assignment variables can be fractional by constraints (7). However, the extent of multiple sourcing or fractional assignment of demands to facilities is bounded in each scenario. In particular, the maximum number of demand nodes that can be fractionally assigned to facilities is less than or equal to $\sum_{j \in J} X_j - 1$ in each scenario. Multiple sourcing may not be overly problematic, if this number is small relative to the total number of demand nodes, $|I|$. In such cases, an approximate solution to the single sourcing problem can often be found for each scenario using the approach suggested by Daskin and Jones (1993). When single sourcing is required and strict optimality is also needed, constraints (7) should be replaced by the obvious integrality constraints

$$Y_{ijs} \in \{0,1\} \qquad \forall i \in I, j \in J, s \in S \tag{10}$$

The imposition of these constraints is likely to increase the difficulty associated with solving the problem considerably.

13.3.2 Implicit Formulation

Model

We next present a formulation of the RFLP in which the random disruptions are modeled implicitly, rather than using explicit scenarios. This formulation is based on the model presented by Snyder and Daskin (2005a). It requires us to make the (rather strong) assumption that the facilities are divided into two sets; the facilities in the first set never fail, while all of the facilities in the second set fail independently with the same probability, q. The first set is called NF (for "non-failable"), while the second is called F (for "failable"). Since the emergency facility never fails, we have $u \in NF$. Note that F and NF constitute a partition of J.

In the implicit formulation of the RFLP, denoted (RFLP2), assignments are made not based on scenarios but based on "assignment levels." In particular, an assignment of customer i to facility j is said to be a "level-r assignment" if there are r open, failable facilities that are closer to i than j is. If $r = 0$, then j is i's "primary" facility—the facility that serves it under normal circumstances—while if $r>0$, j is a "backup" facility. A given customer must be assigned to some facility at every level r from 0 to the number of open facilities, unless it is assigned to some non-failable facility at level $s < r$. We define $Y_{ijr} = 1$ if customer i is assigned to facility j as a level-r assignment.

Since each facility fails with the same probability, we can compute the probability that customer i is served by facility j knowing only the level of i's assignment to j—that is, knowing how many facilities are closer to i but not knowing which facilities those are. This allows a compact formulation of the expected cost. In particular, (RFLP2) is formulated as follows:

$$\text{(RFLP2)} \quad \text{minimize} \sum_{j\in J} f_j X_j + \sum_{i\in I}\sum_{r=0}^{|J|-1} h_i d_{ij}\left[\sum_{j\in NF} q^r Y_{ijr} + \sum_{j\in F} q^r(1-q) Y_{ijr}\right] \tag{11}$$

$$\text{subject to} \sum_{j\in J} Y_{ijr} + \sum_{j\in NF}\sum_{s=0}^{r-1} Y_{ijs} = 1 \;\forall\; i \in I, r = 0,\ldots,|J|-1 \tag{12}$$

$$Y_{ijr} \le X_j \qquad \forall i \in I, j \in J, r = 0,\ldots,|J|-1 \tag{13}$$

$$\sum_{r=0}^{|J|-1} Y_{ijr} \le 1 \qquad \forall i \in I, j \in J \tag{14}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{15}$$

$$Y_{ijr} \ge 0 \qquad \forall i \in I, j \in J, r = 0,\ldots,|J|-1 \tag{16}$$

The objective function (11) minimizes the fixed cost plus the expected transportation cost. The transportation cost term reflects the fact that if customer i is assigned to facility j at level r, then it will be served by j if the r closer facilities fail (which happens with probability q^r) and if j itself does not fail (which happens with probability q if j is failable and with probability 1 if j is non-failable). Constraints (12) stipulate that each customer must be assigned to some facility at each level r, unless the facility is assigned to a non-failable facility at level $s < r$. (By convention, we take $\sum_{s=0}^{r-1} Y_{ijs} = 0$ if $r = 0$.) Constraints (13) prevent an assignment to a facility that has not been opened, while constraints (14) prevent a customer from being assigned to a given facility at more than one level. Constraints (15) and (16) require integrality and non-negativity of the location and assignment variables, respectively. As in the uncapacitated version of (RFLP1), this formulation has an optimal solution in which the assignment variables are binary even though we only require them to be non-negative. Also as in (RFLP1), there exists an optimal facility in which the emergency facility u is open even though we do not explicitly require it. Although assignment levels cannot exceed the number of open facilities, which is not known *a priori*, it is safe to extend the index r to $|J|-1$ in the formulation since each customer is assigned to *some* non-failable facility (possibly u) at some level less than $|J|-1$.

Once the location variables are fixed, it is optimal to assign a customer to its closest open facility at level 0, its second-closest at level 1, and so on, until it is assigned to some non-failable facility (possibly u).

Snyder and Daskin (2005a) propose a Lagrangian relaxation algorithm to solve (RFLP2). They relax constraints (12) to obtain a subproblem that can be solved efficiently to obtain a lower bound for a fixed set of Lagrange multipliers. Upper bounds are obtained by converting the X vector from the lower-bound solution into a feasible solution by assigning customers as described in the previous paragraph. The Lagrange multipliers are updated using subgradient optimization, and the algorithm can be em-

bedded into a branch-and-bound procedure if the bounds produced are not sufficiently tight.

Tradeoff Curve

As discussed above, it is interesting to examine the tradeoff between the UFLP objective and the objective that accounts for failures. Snyder and Daskin (2005a) construct this tradeoff by formulating a multi-objective programming problem with two objectives based on (RFLP2):

$$w_1 = \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ij0} \tag{17}$$

$$w_2 = \sum_{i \in I} \sum_{r=0}^{|J|-1} h_i d_{ij} \left[\sum_{j \in NF} q^r Y_{ijr} + \sum_{j \in F} q^r (1-q) Y_{ijr} \right] \tag{18}$$

Objective w_1 is the classical UFLP objective, while objective w_2 is the objective function from (RFLP2) without the fixed-cost term. We replace the objective function in (RFLP2) with a weighted sum of these two objectives:

$$\text{minimize} \quad \alpha w_1 + (1-\alpha) w_2, \tag{19}$$

where $0 \le \alpha \le 1$. By solving the problem for varying values of α using the weighting method of multi-objective programming (Cohon 1978), we can generate a tradeoff curve consisting entirely of non-dominated solutions. (A solution is *non-dominated* if every other solution is worse than it in at least one of the two objectives.)

The resulting tradeoff curves for the 49-node data set described earlier are depicted in Fig. 13.3 for q = 0.01, 0.05, and 0.10. All facilities are assumed to be failable. The UFLP cost (w_1) is plotted on the x-axis and the failure cost (w_2) is plotted on the y-axis. Each point on a curve represents a different value of α and a different solution.

The solution that is optimal for the classical UFLP (found by solving (RFLP2) with $\alpha = 1$) is the left-most point on each curve. These points are equal on the horizontal axis (since they represent the same solution and hence have the same UFLP cost) but unequal on the vertical axis since they have different failure probabilities and hence different expected failure costs.

Fig. 13.3 suggests that as q decreases, the tradeoff curve shifts. That is, if the firm can somehow reduce the failure probability at its facilities, it can attain a higher level of reliability with the same UFLP cost—or,

equivalently, it can attain the same level of reliability with a lower UFLP cost.

The steepness of the left part of each curve suggests that there are solutions that are much better than the UFLP solution in terms of reliability but not much worse in terms of cost. For example, consider the bottom curve, corresponding to $q = 0.01$. The third point from the left of this curve represents a solution that is 25% better than the UFLP solution in the reliability objective (w_2) but only 7% worse in the UFLP objective (w_1). Similarly, the fifth point is 38% better in w_2 but only 15% worse in w_1. These solutions are depicted in Figs. 13.4 and 13.5.

The number of facilities open in each solution tends to increase as we move rightwards in the curve, since more reliable solutions tend to have more facilities open. The right-most portion of the curve is quite flat, but this portion of the curve is not of much interest because nearly all of the facilities are open in these solutions; they are very reliable but excessively expensive.

We find tradeoff curves with this shape for a wide range of models and data sets, suggesting that large improvements in reliability can often be attained with only small increases in cost.

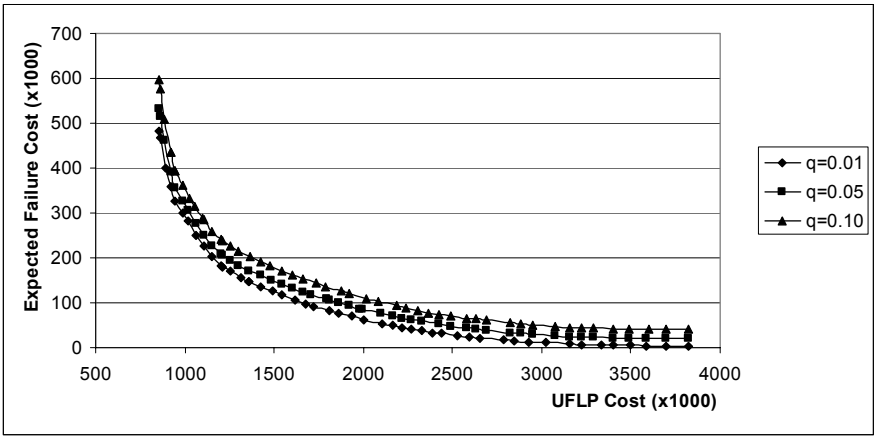


Fig. 13.3. Tradeoff curve for 49-node dataset

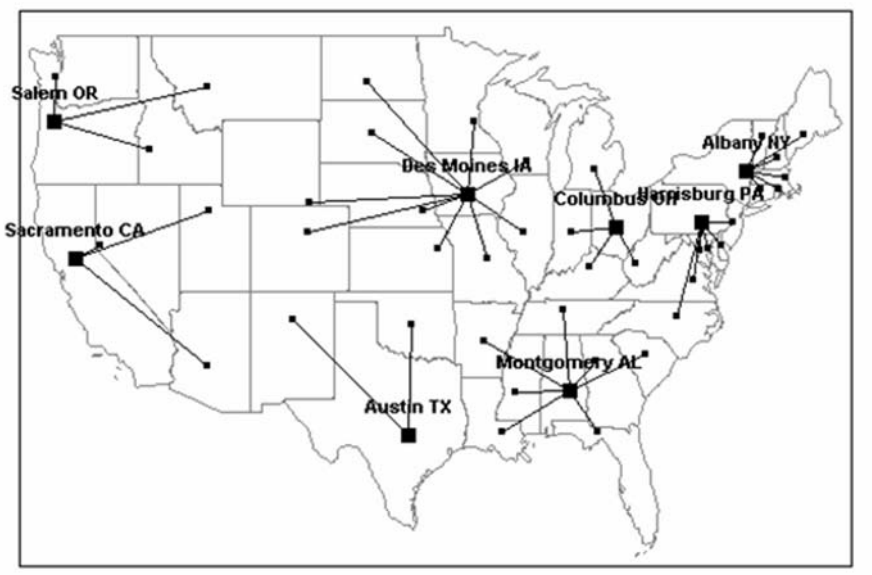


Fig. 13.4. Solution corresponding to third point on $q = 0.01$ tradeoff curve in Fig. 3.

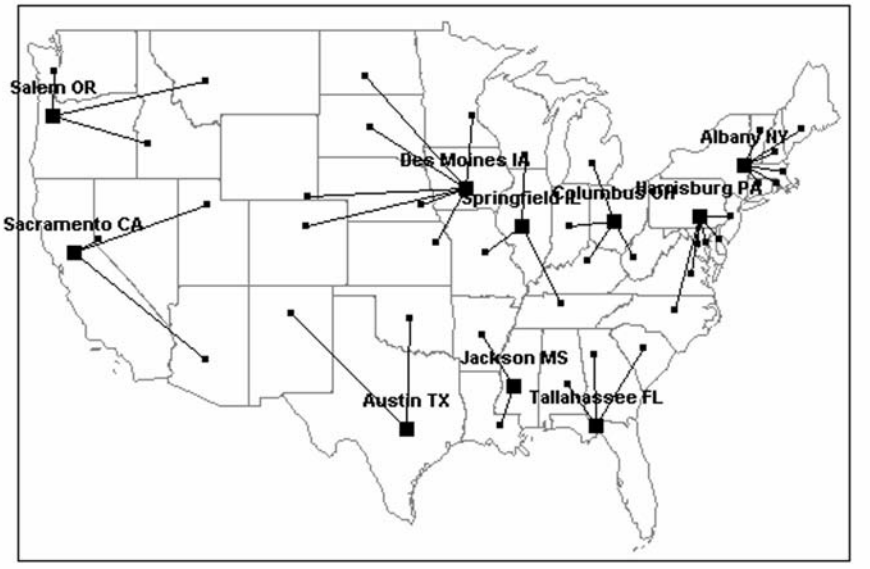


Fig. 13.5. Solution corresponding to fifth point on $q = 0.01$ tradeoff curve in Fig. 13.3.

13.4 Alternate Operating Characteristics and Risk Measures

13.4.1 Introduction

In this section, we outline a number of extensions to the base models defined above. We begin with a variant of the models that allows us to locate two different types of facilities: facilities that are perfectly reliable, or completely hardened against any and all attacks, and facilities that are subject to failure or that are unreliable in some way. The model will determine how many of each type of facility to locate and where they should be. In the second portion of this section, we explore alternative risk measures that also extend the formulations identified above.

13.4.2 Reliable and Unreliable Facilities

In recent years, much attention has focused on the need to harden facilities against attacks. The attacks can be intentional, as in the case of terrorist attacks, or random or unintentional, as in the case of natural disasters. Scaparra and Church (2005a,b) outline defender/interdictor extensions to the traditional *P*-median problem in which *P* facilities *already exist* in a network. A defender can fortify *q* of these facilities against an attack by an interdictor against *r* of the remaining undefended facilities. The objective of the interdictor is to maximize the demand-weighted total distance with demands assigned to the closest non-interdicted facilities, while the defender attempts to minimize this worst-case cost by defending a subset of the facilities. Brown et al. (2005) provide an excellent tutorial on this class of defender/attacker problems.

We adopt a somewhat different approach, first suggested by Daskin (2005) and Daskin et al. (2006). First, we assume that facilities fail randomly. As such, we do not need to model the behavior of an interdictor whose objective is to maximize the damage that he or she inflicts on a network. Second, we do not assume that any facilities exist in the network; rather we formulate the model below based on *de novo* planning with no pre-existing facilities. The model can readily be adapted to the case in which some facilities already exist, through appropriate changes in the fixed costs.

One of two types of facilities can be established at each candidate site *j*. A reliable facility will never fail. Such a facility costs f_j^R at candidate site *j*. Alternatively, we may elect to construct an unreliable facility which can fail with probability *q* but which costs f_j^U. Clearly we require $f_j^U < f_j^R$ for there to be an incentive to locate any unreliable facilities. We define location decision variables X_j^R (and X_j^U) to be 1 if we locate a reliable (or unreliable) facility at candidate site *j* and 0 otherwise.

Similarly, every demand node *i* must be assigned to both a primary facility and a backup facility. The primary assignment will be used if the closest facility has not failed. The backup assignment will be to the closest reliable facility and will be used when the primary facility has failed. Thus, if the primary facility to which a demand node is assigned has failed, the demands at that node are served by the nearest reliable facility, not the nearest facility which has not failed. In this way, the model is a simplification of the base model outlined above. This assignment scheme is chosen primarily for computational reasons. However, during a disruption, real-time information is often limited, and it may be quite reasonable to assume that firms re-assign customers to their nearest reliable facility rather than trying to ascertain whether a closer unreliable facility is operational. We use decision variables Y_{ij}^P and Y_{ij}^B for the primary and backup assignments, respectively.

With this notation, the model becomes:

$$\text{(RFLP3)} \quad \text{minimize} \tag{20}$$
$$\sum_{j\in J} f_j^U X_j^U + \sum_{j\in J} f_j^R X_j^R + (1-q)\sum_{i\in I}\sum_{j\in J} h_i d_{ij} Y_{ij}^P + q\sum_{i\in I}\sum_{j\in J} h_i d_{ij} Y_{ij}^B$$

$$\text{subject to } \sum_{j\in J} Y_{ij}^P = 1 \qquad \forall i \in I \tag{21}$$

$$\sum_{j\in J} Y_{ij}^B = 1 \qquad \forall i \in I \tag{22}$$

$$Y_{ij}^P \le X_j^U + X_j^R \qquad \forall i \in I, j \in J \tag{23}$$

$$Y_{ij}^B \le X_j^R \qquad \forall i \in I, j \in J \tag{24}$$

$$X_j^R + X_j^U \le 1 \qquad \forall j \in J \tag{25}$$

$$\sum_{j\in J} X_j^R \ge 1 \tag{26}$$

$$X_j^R \in \{0,1\} \qquad \forall j \in J \tag{27}$$

$$X_j^U \in \{0,1\} \qquad \forall j \in J \tag{28}$$

$$Y_{ij}^P \ge 0 \qquad \forall i \in I, j \in J \tag{29}$$

$$Y_{ij}^B \ge 0 \qquad \forall i \in I, j \in J \tag{30}$$

The objective function (20) minimizes the total fixed cost for reliable and unreliable facilities as well as the transportation cost for primary and backup assignments. Primary assignments occur with probability $1–q$ for each demand node and backup assignments occur with probability q. If a customer's primary facility is reliable, then its backup assignment will be to the same facility, and the objective function computes the transportation cost to this facility with probability 1. Constraints (21) and (22) require that each demand node be assigned to a primary and backup facility. Constraints (23) state that the primary assignment can only be made to an open

(reliable or unreliable) facility, while constraints (24) state that the backup assignment can only be to a reliable facility. Constraints (25) state that at any candidate site either a reliable or an unreliable facility can be located, but not both. Constraint (26) requires the model to locate at least one reliable facility. Constraints (27) and (28) are standard integrality constraints for the location variables, while constraints (29) and (30) are non-negativity constraints for the primary and backup assignment variables respectively.

Table 13.2. Results from RFLP3 Model for the 49-node dataset

Failure Prob	# Reliable	# Unrel.	Total Cost (x$1,000)	Reliable Sites	Unreliable Sites
0.000	0	13	1,544		CA CO FL IA IL MI MS NY OH OR PA TX VA
0.010	1	12	1,643	PA	CA CO FL IA IL MI MS NY OH OR TX VA
0.030	2	11	1,742	IA PA	CA CO FL IL MI MS NY OH OR TX VA
0.050	3	10	1,805	MS OR PA	CA CO FL IA IL MI NY OH TX VA
0.100	3	9	1,910	IL OR PA	CA CO FL IA MI MS NY OH TX
0.150	4	7	1,992	IL MS OR PA	CA FL IA MI NY OH TX
0.200	4	6	2,046	CA IL MS PA	FL IA NY OH OR TX
0.250	5	4	2,079	AL CA IL PA TX	IA NY OH OR
0.300	5	4	2,107	AL CA IL PA TX	IA NY OH OR
0.350	5	4	2,135	AL CA IL PA TX	IA NY OH OR
0.360	6	3	2,139	AL CA IA OH PA TX	IL NY OR
0.400	6	2	2,153	AL CA IA OH PA TX	NY OR
0.450	6	2	2,168	AL CA IA OH PA TX	NY OR
0.475	6	1	2,174	AL CA IA OH PA TX	NY
0.500	6	0	2,177	AL CA IA OH PA TX	

Constraints (25) and (26) are not strictly needed. In the formulation as stated, there is no incentive to locate both a reliable and an unreliable facility at any candidate site; hence constraints (25) are not needed. Similarly, constraints (26) are implied by the need to provide a backup assignment to a reliable facility for every demand node (constraints 22 and 24). However, in many solution algorithms which relax one or more of the remaining constraints, these constraints are valuable additions as they tighten the relaxed formulation. For example, Daskin (2005) and Daskin et al. (2006) outline an extension of this model that allows the backup distance or cost to differ from the primary distance or cost even for the same demand node/facility pair. This extension requires the incorporation of additional decision variables, additional terms in the objective function and additional

constraints to correct for the case in which a demand node is assigned to a reliable facility as both its primary and backup facility. (This correction is not needed when the primary and backup distances for each demand node/facility pair are the same as is the case in the formulation above.) They outline a Lagrangian solution approach that relaxes constraints (21) and (22) above. Constraints (25) and (26) significantly strengthen the bounds that result from this relaxation.

Table 13.2 shows the results associated with applying the model to the 49-node dataset. In these results, we increased the demand by a factor of 3 compared to the earlier results so that more facilities would be justified in the base case when no facilities are subject to failure. (All costs are in units of $1000.) For all of these runs, the cost of a reliable facility was set to twice the cost of an unreliable facility at each candidate site.

Fig. 13.6 shows the solution when the facility failure probability is 0.05. Fig. 13.7 shows the results for a failure probability of 0.15, while Fig. 13.8 shows the results for a failure probability of 0.25. In all figures, the unreliable sites are shown in *italics*. Some demand nodes are shown with one assignment while others – those whose primary assignment is to an unreliable facility (dashed lines) – are shown with two assignments.

Several observations are worth noting. First, as the probability of a facility failing increases, the number of reliable facilities increases, the number of unreliable facilities decreases and the total cost increases. Second, for moderate values of the failure probability (under 0.05 in this case), the total number of sites does not change from the optimal number found when facilities are not subject to failure, but some facilities are hardened to insure that they do not fail. For larger failure probabilities, the total number of facilities decreases. Third, as the failure probability increases, some facilities will be eliminated completely (e.g., the facility at Richmond, VA which is eliminated once the failure probability gets to 0.10). Some facilities will be converted to reliable facilities as the failure probability increases (e.g., the facility at Harrisburg, PA, which becomes a reliable facility and remains a reliable facility for any failure probability). Other facilities change from unreliable, to reliable, back to unreliable and then back to reliable facilities again as the failure probability increases (e.g., the facility in Des Moines, IA, or the facility in Springfield, IL, which goes from an unreliable site, to a reliable facility and then back to an unreliable site). Finally, some facilities are introduced into the solution as the probability of failure increases (e.g., the facility at Montgomery, AL which enters the solution when the facility failure probability reaches 0.25).

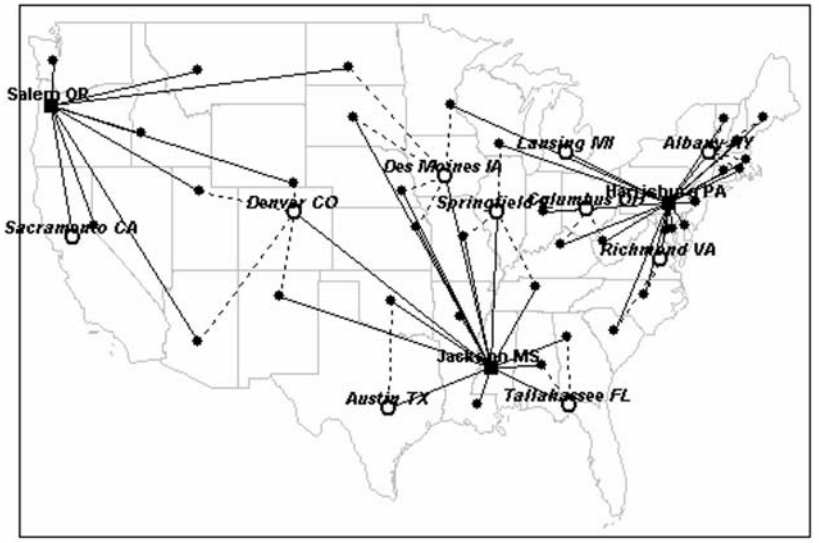


Fig. 13.6. Optimal locations of 3 reliable sites and 10 unreliable sites when failure probability is 0.05

In addition, as the failure probability increases, the expenditure on reliable facilities increases, while the contribution of the fixed facility costs for unreliable facilities decreases. Also, as the failure probability increases, the primary transportation cost increases (as there tend to be fewer facilities overall) but the backup transportation cost decreases (since the number of reliable sites increases with the failure probability). Finally, for failure probabilities exceeding 0.5 in this case, it is not cost-effective to utilize unreliable sites. In fact, an extension of a simple analytic model to incorporate both reliable and unreliable facilities indicates that, under the idealized assumptions of the analytic model (including equal reliable-facility costs of f^R across facilities, and similarly for f^U, and a uniform distribution of demand), unreliable facilities are not employed when the failure probability exceeds $\left(f^R - f^U\right)\Big/ f^R$ (Daskin, 2005; Daskin et al., 2006). While the discrete model whose results are shown above does not require all facility sites to cost the same amount of money, at any candidate site a reliable facility will be twice the cost of an unreliable facility. Thus, loosely speaking, the ratio above will be 0.5 even for the discrete results. As shown in Tab. 1, when the failure probability exceeds 0.5, no unreliable facilities are used.

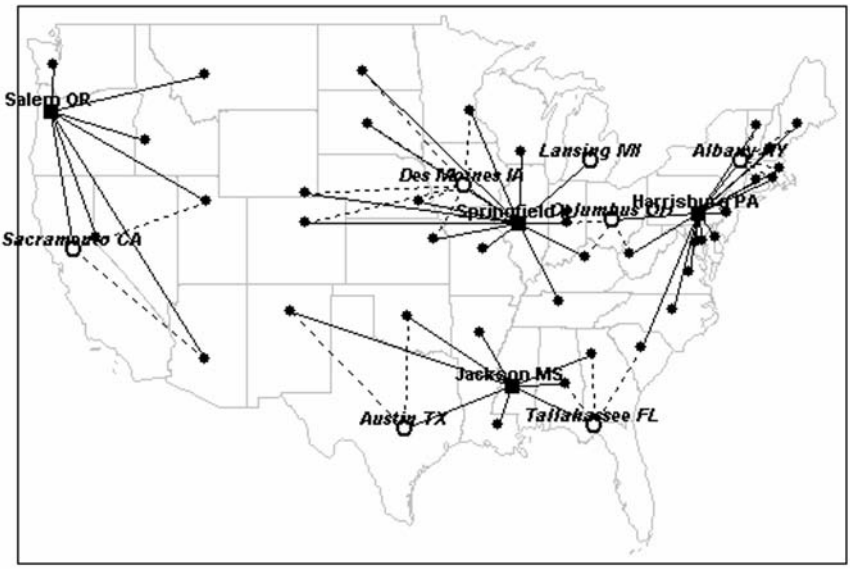


Fig. 13.7. Optimal locations of 4 reliable sites and 7 unreliable sites when failure probability is 0.15

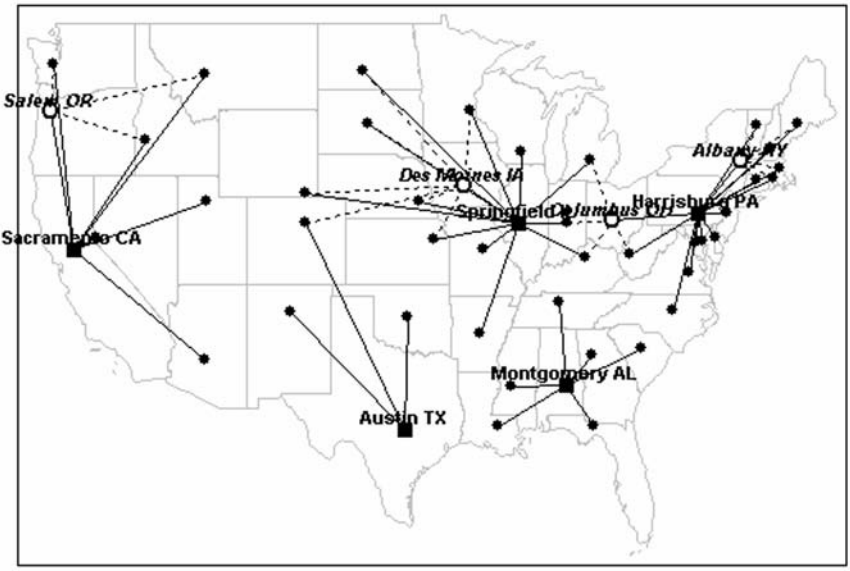


Fig. 13.8. Optimal locations of 5 reliable sites and 4 unreliable sites when failure probability is 0.25

13.4.3 Other Risk Measures

The risk measures discussed so far focus on the average performance of the system when facilities fail. Such models assume that decision makers are risk neutral. In many contexts, decision makers are risk averse: they are concerned not only with the expected performance, but with the potential deviation from it. This may be particularly true when managers are faced with the prospects of losing facilities to natural or man-made disasters. Therefore, in this sub-section, we briefly formulate a number of extensions to the base model that allow decision makers to explore alternate risk measures. In general, these risk measures have all appeared in the literature on facility location under demand uncertainty but have not previously been used for disruption problems.

Minimax Cost Model

The first extension to the base model entails minimizing the worst-case cost in the event of a failure. To do so, we define a new decision variable, U, which is equal to the worst-case fixed plus transportation cost over all scenarios. Objective (31) below minimizes this cost subject to constraint (32), which defines the cost in terms of the total fixed plus demand-weighted transportation cost in each scenario.

$$\text{minimize } U \tag{31}$$

$$\text{subject to} \quad \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ijs} \leq U \quad \forall s \in S \tag{32}$$

$$(4) - (7)$$

This formulation has the advantage of not requiring scenario probabilities as inputs. However, while the expected cost measure defined in (3) is risk neutral, the minimax objective of (31) is extremely risk averse. In fact, the location plan is frequently defined by one (possibly low-probability) scenario, as is often the case in minimax objectives (including the *P*-center model, for example). Such a strong aversion to the worst case often leads to solutions that are quite costly in the non-worst cases. As such, the minimax approach, which places undue emphasis on the worst case, is difficult to justify, just as is the expected value objective of (3), which allows very bad worst-case results. Additional approaches are outlined below.

Mean-Variance

One of the first and most famous objectives considered for optimization under uncertainty is the mean-variance approach. In this model, we minimize a weighted sum of mean cost and the variance of the cost. To define this model, let $z_s(Y)$ be the transportation cost in scenario s if the allocation variables are given by Y. Then the mean-variance model may be formulated as follows

$$\text{minimize} \quad \sum_{j\in J} f_j X_j + \sum_{s\in S} q_s z_s(Y) + \lambda\left[\sum_{s\in S} q_s \left(z_s(Y)\right)^2 - \left(\sum_{s\in S} q_s z_s(Y)\right)^2\right] \tag{33}$$

$$\text{subject to} \quad (4)\text{-}(7)$$

where λ is a weight that is placed on the variance of the transportation costs. The variance places a higher implicit penalty on transportation costs that are significantly larger (and smaller) than the average.

The key problem with this model is that the objective function is highly non-linear. Also, equally penalizing transportation costs that are lower than the average and higher than the average seems somewhat illogical as decision makers are most likely to be concerned with costs that exceed the mean.

Bounding the Cost

One approach to balancing the average cost and the worst-case cost is to minimize one cost while bounding the other. For example, we can minimize the expected cost over all scenarios – objective (3) – while bounding the cost in each scenario. This formulation is shown below.

$$\text{minimize} \quad \sum_{j\in J} f_j X_j + \sum_{s\in S}\sum_{i\in I}\sum_{j\in J} q_s h_i d_{ij} Y_{ijs} \tag{3}$$

$$\text{subject to} \quad \sum_{j\in J} f_j X_j + \sum_{i\in I}\sum_{j\in J} h_i d_{ij} Y_{ijs} \le r \quad \forall s \in S \tag{34}$$

$$(4) - (7)$$

Constraint (34) limits the cost in each scenario, including the fixed facility costs which are common across all scenarios, to a value r. Alternatively, we can simply minimize the uncapacitated fixed charge location problem (UFLP) objective subject to (34) as well as (4)-(7). The UFLP objective is simply:

$$\text{minimize} \quad \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ij} \tag{35}$$

This is equivalent to minimizing the cost in the scenario in which no facilities fail subject to a constraint on the costs incurred when facilities do fail. This approach was proposed by Snyder (2003).

One problem with this approach is that the costs incurred when facilities fail may differ significantly from one scenario to another. Thus, it may make more sense to constrain the costs in scenario *s* relative to the best we could do in scenario *s*, had we known that scenario *s* would occur, rather than relative to some absolute limit *r*. To do so, we define z_s to be the optimal objective function value in scenario *s*. We can then modify (34) to constrain the total cost in scenario *s* to be (1+*r*) times the optimal cost in scenario *s* as shown in constraint (36).

$$\sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ijs} \le (1+r) z_s \quad \forall s \in S \tag{36}$$

Let us define $R_s = \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ijs} - z_s$. R_s is the *absolute regret* in scenario *s*: the absolute difference in total cost between the best we can do in scenario *s* and the best we could have done in scenario *s* had we known that scenario *s* would occur. Similarly, R_s / z_s is the *relative regret*, which represents the percentage difference.

Effectively, (36) constrains the relative regret in each scenario to be no more than *r*. This approach is similar to the "stochastic *p*-robust optimization" approach introduced by Snyder and Daskin (2005b), which minimizes the expected cost in a facility location problem with uncertain demands and costs, subject to a constraint requiring the regret in any scenario to be no more than *p*. Snyder and Daskin argue that stochastic *p*-robustness combines the attractive elements of the min-expected-cost and minimax-cost approaches by optimizing the expected performance while ensuring adequate performance in every scenario. They show that large improvements in robustness (i.e., decreases in worst-case cost) are possible with only small increases in expected cost.

A similar phenomenon is evident in the results of the model in which we minimize the expected cost (3) subject to (36) and (4)–(7). Table 13.4 reports the solutions of this model for various values of *r* for the 49-node data set. For computational reasons, these tests only include scenarios in which zero or one facilities fail. The first column lists *r*, the maximum allowable relative regret. The second column gives the expected cost of the resulting solution (×1000), while the third gives the maximum relative re-

gret of this solution (which must be no greater than r). The fourth column lists the states in which facilities are opened in the solution.

Notice that substantial reductions in regret are possible with only minor increases in expected cost. For example, the second solution has a maximum regret that is 29% smaller than the baseline solution ($r = \infty$) but has only 4% greater expected cost. Similarly, the last solution ($r = 0.25$) has 68% smaller maximum regret but only 6% greater expected cost.

The last row of Table 13.4 corresponds to the optimal solution for the scenario in which the PA facility fails. This scenario is the one that attains the maximum regret for all values of r except ∞. As r decreases, this is the critical scenario, and the solution adjusts to reduce the regret in it. When $r = 0.25$ (corresponding to 25% regret), the solution is quite similar to the optimal solution for that scenario: the two solutions have four facilities in common, two neighboring pairs of facilities (OH / MI and PA / NJ), and only one outlier facility. If we reduce r below 0.209, a second scenario becomes critical, and it is impossible to reduce the regret of both scenarios simultaneously; therefore, the problem becomes infeasible.

This last point highlights one of the main difficulties with models that bound the cost in each scenario. In the other models we have discussed, it is trivial to find a feasible solution. In contrast, as r decreases, it can become quite difficult to find a solution that is feasible with respect to (36). In fact, Snyder and Daskin (2005b) prove that, if the number of scenarios is at least 2, then determining whether a given problem instance is *feasible* is NP-complete. Their result applies to a problem with uncertain demands and costs, but a similar result can be proven for problems with facility failures.

Table 13.4. Solutions to problems with bounded costs

R	Exp. Cost (×1000)	Max Regret	Locations
∞	737	0.649	AL IL NV PA TX
0.5	768	0.462	AL CA IL OR PA TX
0.4	774	0.341	AL CA IN OR PA TX
0.3	776	0.274	AL CA IA MI OR PA TX
0.25	782	0.209	AL CA IA OH OR PA TX
[PA fails]	—	—	AL CA IA MI NJ TX

α-Reliability

In many personal, private sector and public-sector decision contexts, it makes sense to plan not just for the average or expected-value case, or for the worst case, but rather for some eventuality in between these extremes.

For example, the expected outcome of the more than 2 million cosmetic surgical procedures performed in the U.S. in 2004 (ASAPS 2004) was an improvement in the patient's appearance. The worst-case outcome undoubtedly was death in a small percentage of the cases. While most people who have such elective surgery expect the best possible outcome, it is prudent to plan for adverse results as well. For example, it is wise to have an up-to-date will as well as a living will and health care proxy before undergoing any surgical procedure. Similarly, in the design of public facilities such as airports, we clearly do not plan just for the average volume, but we also do not size airports for the peak demands associated with the Thanksgiving weekend. Airport capacity is based on values that are intermediate between the average and maximum daily demand levels.

In a similar manner, we can plan against an endogenously determined subset of the scenarios whose combined probability is at least α. One variant of this approach would minimize the maximum regret over all such scenarios, ignoring the regret in the remaining scenarios. To do this, we define a new variable W_s to be 1 if scenario s is in the "reliability set" against which we are planning and 0 otherwise. (Note that the term *reliability* is used in a different context in this model and refers to an endogenously determined set of scenarios against which the model is planning.) We also define R to be the maximum regret over all scenarios in the reliability set. Finally, we let M be a large number, larger than any possible scenario regret. With this notation, the problem can be formulated as:

$$\text{minimize } R \tag{39}$$

$$\text{subject to } \sum_{s \in S} q_s W_s \geq \alpha \tag{40}$$

$$\sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ijs} - z_s - M(1 - W_s) \leq R \quad \forall s \in S \tag{41}$$

(4)-(7)

The objective function (39) minimizes the maximum regret R over the scenarios in the reliability set. Constraint (40) requires the reliability set over which the minimization is performed to have a probability of at least α. Constraint (41) defines the maximum regret in terms of the scenario-based regrets, but excludes scenarios that are not part of the reliability set. This model was first proposed by Daskin et al. (1997), though scenarios in the original model referred to uncertainty in demand rather than in supply.

One problem with the α-reliable minimax regret model above is that it ignores the regret associated with scenarios that are not part of the reliability set. Chen et al. (2005) have proposed the α-reliable mean excess regret model, which, when applied to the problems at hand results in the following formulation:

$$\text{minimize } \zeta + \frac{1}{1-\alpha} \sum_{s \in S} q_s U_s \tag{42}$$

$$\text{subject to } U_s \geq R_s - \zeta \quad \forall s \in S \tag{43}$$

$$\sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ijs} - z_s \leq R_s \quad \forall s \in S \tag{44}$$

(4)-(7)

In this model, we can think of ζ as the regret contribution of every scenario. A fraction of the scenarios (approximately equal to α) will have regret values that exceed this endogenously determined value. The objective function (42) minimizes the sum of ζ and the expected regret in excess of this value. Constraint (43) defines the excess regret in scenario s as the amount by which the regret in scenario s exceeds the nominal value ζ. Constraint (44) defines the regret in scenario s in terms of the compromise locations, the scenario-specific demand assignments and the optimal objective function for scenario s, z_s.

Although the α-reliable minimax regret and α-reliable mean excess regret models were originally formulated for problems with demand uncertainty, they can be applied equally well to problems with facility failures.

Chance-Constrained Approach

Finally, we note that the α-reliable minimax regret model is similar to a chance-constrained model. For example, we can minimize the expected cost over all scenarios – objective (3) – subject to a constraint that the sum of the probabilities associated with scenarios in which the cost exceeds some value C_{target} is less than or equal to β, a user-specified value. Let us define the decision variable T_s to be 1 if the cost in scenario *s* exceeds C_{target} and 0 otherwise. A chance-constrained model can now be formulated as follows:

$$\text{minimize} \sum_{j \in J} f_j X_j + \sum_{s \in S} \sum_{i \in I} \sum_{j \in J} q_s h_i d_{ij} Y_{ijs} \tag{3}$$

$$\text{subject to} \sum_{s \in S} q_s T_s \leq \beta \tag{45}$$

$$\sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ijs} - C_{\text{target}} \leq M T_s \quad \forall s \in S \tag{46}$$

$$(4)\text{-}(7)$$

The objective minimizes the expected cost over all scenarios. Constraint (45) states that the sum of the probabilities of all scenarios with cost greater than C_{target} must be less than or equal to β. Constraint (46) links the location and allocation variables which define the cost in scenario *s* to the variables T_s. If the scenario-specific cost exceeds C_{target}, meaning that the left-hand side of (46) is positive, then T_s must be 1; otherwise T_s may be 0. Again, *M* is a sufficiently large value so that constraint (46) will not be binding whenever $T_s = 1$.

13.5 Conclusions

Supply chain planners face a significant amount of uncertainty, particularly during the strategic planning phase. Facility location decisions are very expensive to change, so planners must take uncertainty into account when choose facility locations. In this chapter, we have illustrated the broad range of strategies that decision makers might take for approaching risk in facility location models with supply disruptions. A planner may choose one or more of these approaches based on his or her level of risk aversion, the type of disruptions that are of greatest concern, the flexibility of each measure to fine-tune parameters and add side constraints, the computational difficulty with which each model can be solved, and other factors.

One key insight that comes from many of the models we have discussed is that it is often relatively inexpensive to "buy" reliability—that is, if decision makers are willing to sacrifice just a bit in the objectives they are used to considering, they can gain significant improvements in other objectives, including reliability.

The models discussed in this chapter by no means represent an endpoint for research on facility location with disruptions. Several important issues remain to be addressed. One is computational: Many of these models are simply too difficult to solve, for reasonably sized instances, using off-the-shelf IP solvers. Rather, special-purpose algorithms, such as those proposed by Snyder and Daskin (2005a,b) and others, must be developed to solve these problems.

Another important direction for future research involves capturing other types of supply chain decisions in a unified model. A number of models attempt to incorporate tactical decisions, such as inventory and vehicle routing, into the facility location decision. These models tend to offer a substantial improvement over a sequential optimization approach in which facility locations are chosen first, and then tactical decisions are made while keeping the strategic decisions fixed. A natural next step is to consider facility failures in these models. For example, Jeon, Snyder, and Shen (2006) consider facility failures in the context of the joint location-inventory model first proposed by Daskin, Coullard, and Shen (2002) and Shen, Coullard, and Daskin (2003).

A third avenue for future research involves multi-echelon facility location and network design problems with disruptions. Such models might be based on the seminal distribution network design problem of Geoffrion and Graves (1974). In the multi-echelon case, a key question is how to model the "cascading" effect of disruptions, as failures at one echelon lead

to failures downstream, either explicitly (because of geographical proximity of the facilities, for example) or implicitly (as downstream facilities become starved for raw materials during a disruption). We hope that this chapter will help to spark future research on these and other related topics.

References

ASAPS (American Society for Aesthetic Plastic Surgery) (2004), Cosmetic Surgery Quick Facts: Highlights of the ASAPS 2004 Statistics on Cosmetic Surgery, http://www.cosmeticplasticsurgerystatistics.com/statistics.html, accessed March 2006

Balinski ML (1965) Integer programming: Methods, uses, computation. Management Science 12:253-313

Barrionuevo A, Deutsch CH (2005) A distribution system brought to its knees. New York Times, Sep. 1, 2005, p. C1

Berman O, Krass D (2002) Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher HW (eds) Facility location: Applications and theory. Springer-Verlag, New York, chapter 11.

Berman O, Krass D, Menezes MBC (2005a) Facility reliability issues in network p-median problems: Strategic centralization and co-location effects. Forthcoming in Operations Research

Berman O, Krass D, Menezes MBC (2005b) MiniSum with imperfect information: Trading off quantity for reliability of locations. Working paper, Rotman School of Management, University of Toronto, Toronto, Canada

Berman O, LeBlanc B (1984) Location-relocation of mobile facilities on a stochastic network. Transportation Science 18:315-330

Berman O, Larson RC, Chiu SS (1985) Optimal server location on a network operating as an M/G/1 queue. Operations Research 33:746-771

Birge JR, Louveaux F (1997) Introduction to stochastic programming. Springer-Verlag, New York.

Brown GG, Carlyle WM, Salmerón J, Wood K (2005) Analyzing the vulnerability of critical infrastructure to attach and planning defenses. In: Greenberg HJ (ed) TutORials in operations research. INFORMS, Baltimore, pp. 102-123

Chen G, Daskin MS, Shen ZJ, Uryasev S (2003) The α-reliable mean-excess regret model for stochastic facility location. Working paper, University of Florida, Department of Industrial and Systems Engineering

Church RL, Scaparra MP (2005) Protecting critical assets: The *r*-interdiction median problem with fortification. Working paper 79, Kent Business School, Canterbury, England

Cohon JL (1978) Multiobjective programming and planning. Academic Press, New York

Colbourn C (1987) The combinatorics of network reliability. Oxford University Press, New York

Daskin MS (1982) Application of an expected covering model to EMS system design. Decision Sciences 13:416-439

Daskin MS (1983) A maximum expected covering location model: Formulation, properties and heuristic solution. Transportation Science 17:48-70

Daskin MS (1995) Network and discrete location: Models, algorithms, and applications. Wiley, New York

Daskin MS, Chopra SL, Lim M (2005) To harden or not to harden. Presentation, University of Kent Business School, Canterbury, England, October 2005

Daskin MS, Chopra SL, Lim M (2006) Integrated models of reliable and unreliable facility location problems. Working paper, Northwestern University, Evanston, IL, Department of Industrial Engineering and Management Sciences

Daskin MS, Coullard CR, Shen ZJM (2002) An inventory-location model: Formulation, solution algorithm and computational results. Annals of Operations Research 110:83-106

Daskin MS, Hesse SM, ReVelle CS (1997) α-reliable p-minimax regret: A new model for strategic facility location modeling. Location Science 5:227-246

Daskin MS, Hogan K, ReVelle C (1988) Integration of multiple, excess, backup, and expected covering models. Environment and Planning B 15:15-35

Daskin MS, Jones PC (1993) A new approach to solving applied location/allocation problems. Microcomputers in Civil Engineering 8:409-421

Daskin MS, Snyder LV, Berger RT (2005) Facility location in supply chain design. In: Langevin A, Riopel D (eds) Logistics systems: Design and operation. Springer, New York, pp. 39-66

Drezner Z (ed) (1995) Facility location: A survey of applications and methods. Springer-Verlag, New York

Drezner Z, Hamacher HW (eds) (2002) Facility location: Applications and theory. Springer-Verlag, New York

Fox J (2005) A meditation on risk. Fortune 152:50-62

Garey MR, Johnson DS (1979) Computers and intractability: A guide to the theory of NP-completeness. W. H. Freeman and Company, New York

Geoffrion AM, Graves GW (1974) Multicommodity distribution system design by Benders decomposition. Management Science 20:822-844

Greenhouse S (2002) Both sides see gains in deal to end port labor dispute. New York Times, Nov. 25, 2002, p. A14

Higle JL (2005) Stochastic programming: Optimization when uncertainty matters. In: Greenberg HJ (ed) TutORials in operations research. INFORMS, Baltimore, pp. 30-53

Jeon HM, Snyder LV, Shen ZJM (2006) Joint location-inventory optimization with unreliable facilities. Working paper, Lehigh University, Department of Industrial and Systems Engineering

Kleywegt AJ, Shapiro A, Homem-de Mello T (2001) The sample average approximation method for stochastic discrete optimization. SIAM Journal on Optimization 12:479-502

Larson RC (1974) A hypercube queueing model for facility location and redistricting in urban emergency services. Computers and Operations Research 1:67-95

Leonard D (2005) “The only lifeline was the Wal-Mart”. Fortune 152:74-80

Linderoth J, Shaprio A, Wright S (2002) The empirical behavior of sampling methods fo stochastic programming. Working paper, University of Wisconsin-Madison, Computer Sciences Department

Marquis C, McNeil DG (2001) Meat from Europe is banned by U.S. as illness spreads. New York Times, Mar. 14, 2001, p. A1

Mouawad J (2005) Katrina's shock to the system. New York Times, Sep. 4, 2005, p. 3.1

Owen SH, Daskin MS (1998) Strategic facility location: A review. European Journal of Operational Research 111:423-447

Pollack A (2004) U.S. will miss half its supply of flu vaccine. New York Times, Oct. 6, 2004

Reuters (2005) Lumber, coffee prices soar in Katrina's wake. Reuters, Sep. 1, 2005

ReVelle C, Hogan K (1989) The maximum availability location problem. Transportation Science 23:192-200

Scaparra MP, Church RL (2005) An optimal approach for the interdiction median problem with fortification. Working paper #78, Kent Business School, Canterbury, England

Scaparra MP, Church RL (2006) A bilevel mixed integer program for critical infrastructure protection planning. Working paper, Kent Business School, Canterbury, England

Shen ZJM, Coullard CR, Daskin MS (2003) A joint location-inventory model. Transportation Science 37:40-55

Shier D (1991) Network reliability and algebraic structures. Clarendon Press, Oxford

Shooman ML (2002) Reliability of computer systems and networks: Fault tolerance, analysis, and design. John Wiley & Sons, New York

Snyder LV (2003) Supply chain robustness and reliability: Models and algorithms. PhD thesis, Northwestern University, Evanston, IL, Dept. of Industrial Engineering and Management Sciences

Snyder LV (2006) Facility location under uncertainty: A review. IIE Transactions 38:537-554

Snyder LV, Daskin MS (2005a) Reliability models for facility location: The expected failure cost case. Transportation Science 39:400-416

Snyder LV, Daskin MS (2005b) Stochastic *p*-robust location problems. Forthcoming in IIE Transactions

Snyder LV, Scaparra MP, Daskin ML, Church RC (2006) Planning for disruptions in supply chain networks. In: Greenberg HJ (ed) TutORials in operations research. INFORMS, Baltimore, forthcoming

Snyder LV, Ülker NŞ (2005) A model for locating capacitated, unreliable facilities. IERC Conference, Atlanta, GA, May 2005

14 Moving from Protection to Resiliency: A Path to Securing Critical Infrastructure

Laurie Anne Schintler, Sean Gorman, Rajendra Kulkarni and Roger Stough

School of Public Policy, George Mason University, USA, Emails: lschintl@gmu.edu, sgorman1@gmu.edu, rkulkarn@gmu.edu, rstough@gmu.edu

14.1 Introduction

The events of 9/11 brought renewed focus to critical infrastructure, but the security of infrastructure has been and continues to be an issue outside the scope of any one event or country. Oil pipeline attacks in Iraq, massive blackouts in Italy, the United States, and Russia, submarine cable failures in the Atlantic, accidental and intentional failures of infrastructure are an increasing and complex problem. The issue of infrastructure security is a global problem both is applicability and connectivity. All nations are dependent on infrastructure and many of these infrastructures cross international borders and some span the globe. A problem facing all nations is that they have the responsibility for securing infrastructure but critical aspects are owned by the private sector. This though is only one of many problems facing infrastructure security: 1) infrastructures are interdependent on each others reliability 2) infrastructures are large, dynamically unsynchronized, and complex 3) sharing information about infrastructure vulnerabilities is severely hampered by fears of regulation and competition. Along with these direct obstacles there are larger economic forces that complicate the issue. The markets driving infrastructure are geared towards maximizing efficiency to increase profit and not maximizing protection, which can result in public vulnerabilities.

The way in which the aforementioned factors have contributed to security vulnerabilities is well-documented in the telecommuncations sector,

for example (Albert et al 2000, Callaway et al 2001, Cohen et al 2001). It has also between pointed out by the Federal Communication Commission's (FCC) Network Reliability and Interoperability Counsel (NRIC) that:

"Technical and market forces have reduced reserve capacity and the number of geographically diverse, redundant routings in the Public Telecommunications Network (PTN). Failure of a single link can now have serious repercussions (NRIC 1997)."

Interdependencies between telecommunications and the financial services sector have resulted in additional vulnerabilities. The impact of advanced telecommunications has been particularly profound on the structure of financial services (Power 2002, Townsend 2001, OTA 1995, Warf 1989). Perhaps, most significant has been the development of global market place for financial services, made possible by international information networks combines with the deregulation of financial markets (OTA 1995). The deregulation of both telecommunication and financial services has led to agglomeration of both in key global cities resulting in an interdependent collocation of wires and dollars (Townsend 2001, Power 2002, Longcore and Rees 1996, Obrien 1992). While the agglomeration of activities has introduced increasing efficiencies in both sectors it has resulted in negative externalities. The greatest of which was illustrated by the events of September 11th, security. The fall of the World Trade Center towers wreaked havoc on financial markets closing them for six days, largely as a result of the need to provision the 2 million data circuits (15,000 for the NYSE alone) and 19 sonet rings destroyed in the collapse (DOE 2003). The many problems in the financial sector resulting from the attack has brought new attention to the vulnerability of the US and global financial sector.

The result is an impasse between government and the private sector. Most infrastructure providers do not have financial resources to protect their infrastructure from low probability events like terrorism and natural disasters. Thus business models and profits are built on optimizing networks to withstand high probability single failures and not low probability multi-failures, like terrorism or natural disasters. To build or augment an infrastructure to the level of protection desired by the government would put that infrastructure at a competitive disadvantage. The money spent on protection would increase the cost of the service well above the rate required to be competitive with the market. The result is a prisoner's dilemma where no provider wants to invest because unless all infrastructure owners invest they will be at a disadvantage. As a result infrastructure

owners seldom want to share information about vulnerabilities because in doing so there is the chance they will have to invest to remedy those problems, and then be at a competitive disadvantage to the rest of the market that does not disclose vulnerabilities. When one adds to this driving market forces, the fear of regulation, and a disclosure of data to competitors, the result is a quagmire with little hope of progress.

There are possible solutions to the quagmire, but the interface between government and business must be speaking the same language. In the business world the words are not protection and vulnerability, but continuity and resiliency. Even in a world of efficiency and profit maximization there is a requirement for the private sector to be up and available for business, often 24 hours a day 7 days a week. As a result infrastructure needs be equally reliable and there is a booming business in business continuity planning and management. The emphasis in this business lexicon is not protection and vulnerability. Business continuity plans do not include jersey barriers, armed troops, and armored vehicles. In short there is a business case to be made for resilient continuity but not one for protection. If the impasse between government and the private sector is to be surpassed there needs to be a move from critical infrastructure protection to critical infrastructure resiliency. To be successful though there needs to be standards and benchmarks by which to measure continuity by quantifying resiliency and creating metrics to measure the cost effectiveness of investments. Then the private sector, and relevant government stakeholders can make a case that investments in resiliency create a competitive advantage and increases shareholder value.

This paper presents a framework for measuring and visualizing the resiliency of networks. The framework builds on complex network theory and provides a set of metrics that can be input to any fiscal analysis model to identify cost-effective or optimal strategies for enhancing the resiliency of a network. It is also embedded in a Geographic Information System (GIS) to aid in the selection of potential mitigation strategies and to visualize the results of the analysis. The next section of this paper, we provide a high-level overview of the analytical framework proposed. Following this, is a review of the literature on complex network theory and a discussion of how this theory has been applied previously to explore network resiliency. The details of the framework proposed in this paper are then discussed and the method is applied to a portion of the Washington, D.C. electric transmission line network to demonstrate the feasibility and practicality of the approach. The paper concludes with some caveats of the approach and directions for future research.

14.2 Evaluating the Resiliency of Critical Infrastructure: A General Framework

The overall framework for generating information which can be used to identify cost-effective or optimal strategies for improving the resiliency of a network is comprised of four steps: infrastructure assessment, verification and consequence identification.

Infrastructure Assessment

One of the significant obstacles in dealing with critical infrastructure is assessing and setting baselines for such large complex sprawling networks. Further, infrastructures are often interdependent and dynamic. Fortunately there has been considerable work done on methods for quantifying critical infrastructure. Infrastructures can be assessed based on several factors a few of which include:

- Density – how much infrastructure is there in any discrete location – i.e. 15 fiber optic conduits, 3 electric transmission lines, 2 gas pipelines.
- Capacity – how much volume, flow, or traffic are the infrastructures in any discrete location able to handle – i.e. the fiber lines have a 10 Gbps[1] capacity, the electric transmission lines are 720 Kv, and the gas pipeline are 42 inches in diameter.
- Bottleneck identification – algorithmic approaches to identify areas with high amounts of capacity but little diversity to route it.
- Structural analysis – another algorithmic approach that calculates all possible paths across an infrastructure and finds those discrete locations that are most frequently used in routing.
- Weighted structural analysis – expands the all possible path analysis to include to identify those locations are frequently used in routing and have low levels of capacity, or alternative routing paths in the event of failures that could be under capacitated.
- Interdependency –
    - Colocation – two or more infrastructures are located in the same discrete location.

[1] Gigabyte per second (OC-192)

    - Structural – the most frequently utilized routing paths of two or more infrastructures are located in the same discrete location.
    - Functional – the loss of one infrastructure will cause failures in a dependent infrastructure – i.e. the loss of electric power causes traffic light failures resulting in cascading traffic congestion, hampering emergency response.
- Cost – creating a baseline figure of cost for the infrastructure in its current configuration – i.e. the cost of leasing fiber per month from a network provider.

These are but a few of the possible approaches to assessing infrastructure, but they provide a first cut and the basic aspects of infrastructure that are important to understand. Each approach provides a list of discrete locations and assets that could be critical to the operation of one or multiple infrastructures. To know if the assessment created the correct output the analysis needs to be verified.

Verification

There a multitude of ways to assess infrastructure and identify potential vulnerabilities, and there needs to be a means to identify which approach works best in each environment through a verification process. One means of verification is through failure simulation. Once infrastructure has been assessed and the most critical infrastructure components identified and ranked a failure of each component can be simulated. After the failure, the impact can be charted and subsequently compared to other components to verify their criticality. Would the failure of a location with the highest density of infrastructure cause more impact than an area with the highest capacity, or would a failure at a bottleneck cause the greatest repercussions to continuity. Failure simulation provides a means to verify the criticality of any of these scenarios to the continuity of the infrastructure. Once baseline verification has been performed a combination of assessment methods can be investigated. For instance the greatest impact to continuity could come from the most frequently used routing path that contains a high density of three different infrastructures.

A second aspect that needs to be considered in a verification process is after an initial failure the structure of the network changes. What was once the second most critical asset in the network may have changed. To determine if it has or not a combinatorial optimization needs to be run where

after the first failure has been incurred all possible second most critical assets need to be tested to determine which has the greatest impact on the infrastructure's continuity. In the best case scenario real time analysis can be performed to react to failures and determine how to best allocate resources in the network, but proactive analysis before events is still critical to ensure continuity.

Consequence

An integral part of both assessment and verification is determining what the consequences of a failure are in terms of resiliency. Consequences can be calculated through a variety of methods and ultimately are specific to an individual scenario. Metrics for examining the impacts of failures and mitigation strategies on resiliency of a network can include those that reflect the network's structural properties, its level of service defined by traffic or congestion levels or on its users. Some of the broad areas into which consequence or impacts on users can be categorized include:

- Population affected – how many people will be affected by a failure or lack on continuity in an infrastructure – i.e. after a transmission line failure and subsequent blackout how many people will be without power.
- Businesses affected – how many business locations will be affected in a failure scenario for aggregation purposes these consequences can be grouped by SIC or NAICs codes.
- Interdependent infrastructures affected – what infrastructures with dependencies to a failed infrastructure will be impacted by an event – i.e. a transmission line failure causes traffic signals to loose power causing cascading gridlock in transportation infrastructure.

These are just three broad categories under which consequence could be grouped. For specific critical sectors consequence can be more narrowly defined and quantified.

14.3 Literature Review: Complex Network Theory and Resiliency Analysis

There is a large body of literature on complex network theory and the use of this theory to analyze the properties of networks and their resiliency. Spatial applications of graph theory have a long lineage in both geography and regional science. Garrison (1960) did in-depth network analysis on the interstate highway system, analyzing the importance of nodes and links on location and development. This same vein of research was greatly expanded through Garrison's student Kansky (1963) and later with the work of Chorley and Haggett (1969). In addition, Nyusten and Dacey (1968) and later Taffee and Gauthier (1973) expanded this research, applying network analysis to telephone networks and general infrastructure. This tradition of network analysis was picked up again by geographers to begin to analyze the Internet's network of networks. Wheeler and O'Kelly (1999) examined the basic graph measures of several domestic US providers and analysis of city connectivity of the aggregated providers. Gorman and Malecki (2000) investigated the network topologies of several firms and how graph theoretic measures could be used to investigate competitive advantage and the nature of interconnection between networks. Later studies have looked at the structure of networks and city connectivity as a time series finding large changes in bandwidth capacity (Malecki, 2002; Townsend, 2001), but little change in graph measures of connectivity (O'Kelly and Grubesic, 2002). While connectivity indices have changed little over time the overall structure of the network has. Gorman and Kulkarni (2004) found that the aggregated US backbone network has become increasingly self-organized from 1997 to 2000 creating a more efficient but more sparsely connected network. This research confirmed at a spatial level of analysis what was being found at a topological level in the study of complex networks.

In studies of large complex network of thousands and millions of nodes physicists and computer scientists have begun to incorporate spatial dimensions to their work. Work by Yook, Jeong, and Barabasi (2001) has examined the role of linear distance in complex networks. They found that the spatial layout of the global Internet router network formed a fractal set, determined by population density patterns around the globe (Yook et al 2001). A similar study at Boston University found the same effect when population was controlled for with the per capita GDP of regions (Lakhina et al 2002). Barthelemy (2003) found that in spatial networks with scale free properties long distance links connect predominantly to hubs. Further,

if the total length in a network is fixed, the optimal network which minimizes both the total length and the diameter lies in between the scale-free and spatial networks (Barthelemy 2003).

The analysis of the resiliency of networks also has a long history of analysis with applications in fields such as landscape ecology (Urban and Keitt 2001). In addition to several discipline specific approaches there has been considerable recent work on the resiliency of general complex networks. A widely discussed work by Albert et. al. (2000) found that complex networks[2] were robust to random failures but vulnerable to targeted attack. The research illustrated that when nodes with a significant percentage of the networks connections are targeted for attack the network degrades rapidly leading to catastrophic failures and network balkanization.

The initial work by Albert et. al. was quickly followed by several other approaches to vulnerability of large complex networks. Callaway et al (2000) modeled network robustness and fragility as a percolation and Cohen et al (2001) using similar percolation models, both findings reinforcing the fragile-robust dichotomy discovered by Albert et. al. (2000). The research has not been without criticism, and some computer scientists and engineers have argued the network topology models generated in these studies are not accurate (Chen et al. 2001). In fact the same heavy tail connectivity distribution can result from a wide variety of network topologies, some more and less resilient than others (Schintler et al. 2005).

Combining the themes of spatial network analysis and resiliency dates back over thirty years (Haggett and Chorley 1969). Analysis of the resiliency of spatial networks has again been picked up largely in relation to the analysis of critical infrastructure networks. Utilizing a model of node connectivity and path availability Grubesic et al (2003) found that the disconnection of a major hub city could cause the disconnection of peripheral cities from the network. Building upon the complex network literature Gorman et. al. (2004) found that incorporating spatial variables into algorithms, such as global connections between cities and Euclidean distance, to determine the criticality of nodes in the network was more effective than the binary connectivity measures used in the previous studies citied.

[2] Specifically scale free complex networks with power law connectivity distributions.

14.4 A GIS Approach for Assessing the Resiliency of Networks

This section discusses how the resiliency of networks can be assessed using GIS and analytical techniques based on complex network theory. A GIS is used in the process for three reasons: 1. to define the topology of the network, 2. to aid in the identification of potential mitigation strategies for that network and 3. to visualize the result of the analysis. The process uses three types of software: a GIS and spatial analysis software, a programming environment and a network analysis tool.

The process begins with an analysis of the base infrastructure network – i.e., the network without any mitigation. Any type of infrastructure can be analyzed as long as it can be described as a set of links and nodes and a polyline theme file for the network is available. Some examples of networked infrastructure that can be analyzed include the electric power grid, interstate highway system, pipeline network, fiber optic cable and the rail system. The analysis can also be performed at any geographic level.

The first step is to overlay a grid onto the polyline theme for the base network. The size of the grid in terms of rows and columns needs to be defined and this should be done such that each cell has no more than one segment. This is necessary to adequately reflect the detail of the network topology. Each cell containing part of the network is assigned a unique number and this information is then used to create an edge list that describes how the network is structured. A computer code can be written in a program to carry out this function efficiently.

The edge list that is generated is next input into a network analysis program to measure the criticality of each of the vertices in the network[3]. The criticality of a vertex here is defined in terms of the total number of nodal pair combinations in the network that have shortest paths that route through that segment. It is an indicator of how structurally important that link is to the network. In social network theory, this measure is referred to as betweenness centrality (Freeman LC 1979). Other disciplines use similar measures of criticality – e.g., load in physics (Goh et al 2001) and accessibility in transportation (Garrison 1960).

In the next step, the betweenness scores contained in this file are joined back to the cells in the grid overlay that was previously generated. The cells weighted by betweenness scores are is used to generate a density map

[3] The vertices are essentially the centroids of the grid cells that contain points from the network and they define the A and B nodes of each link in the network being analyzed

that provides a visual of the overall resiliency of the network and the location of bottlenecks where there are relatively high betweenness scores and few alternative routes.

A similar process can be used to analyze how different mitigation strategies change the resiliency of the base network. Before carrying out the analysis, the potential candidates for mitigation need to be identified (Step 7). One approach for doing this is to look for existing rights-of-ways that are used by other infrastructures. This can be done by overlaying the polyline themes for other infrastructures on to the base network polyline theme and then by looking for locations where feasible connections could be added to the base network along existing rights-of-ways. The layer for the density map of betweenness scores can further aid in this process by highlighting where additional capacity may most effectively provide alternative routing around high usage bottlenecks in the network. Once mitigation strategies have been identified, the base network polyline them is updated by adding in the new segments. Once this is complete then the process circles back to the first step that involves the grid overlay and ultimately generates a new density map showing the impact of the mitigation strategy on resiliency.

The betweenness scores generated in Step 5 of the process can also be used in another way to measure the overall resiliency of a particular network and to provide some metrics by which the resiliency of different networks can be compared. The approach draws from complex network theory and the idea that a network is less resilient if there are a few highly critical links or nodes, whose removal from the network would have severe negative consequences on the connectivity of the network and possibly balkanization. Based on this hypothesis, a less resilient network would show a skewed distribution of betweenness scores where there are few highly critical links based on their betweenness scores and multiple less critical links. This would most likely be represented as either a power-law or exponential distribution.

To carry out the analysis, the betweenness scores for each of the networks being analyzed first need to be normalized by their maximum values. Next, each set is sorted separately in descending order, the distribution of values in each series graphed and equations fit to each of these lines. Some experimentation may be required to determine whether a power law or exponential distribution is most appropriate for each of the lines or that neither is appropriate and in this case the network would appear to be relatively resilient. The functional form of the equations fit to each of the distributions should be consistent to allow for valid comparisons across networks. The impacts of different mitigation strategies on resiliency can be explored by looking at the change in slope of the distribu-

tion or by comparing other statistics such as the mean, median and standard deviation of the betweenness scores. These metrics can then be input into a financial evaluation method to identify investment strategies. For example, the percentage improvement in resiliency measured by the change in the coefficient in the estimated model distribution can be used in a cost-effective ratio. These techniques are illustrated using a portion of the Washington, D.C. area power grid.

14.5 Evaluating Network Resiliency: An Empirical Example

This section implements the process described in Section 14.4 and analyzes the resiliency of a portion of the Washington, D.C. area electric transmission grid and a potential strategy to enhance the resiliency of that network. Figure 14.1 is a density map of the normalized betweenness scores for the base network. The program UCINET 6.26 (Borgatti et. al. 2003) was used to calculate these scores and ArcView 3.3 to generate the point theme for the network, the grid overlay for the edge list and to generate the density maps. A code was written in Matlab 5.5 to convert the grid overlay to an edgelist.

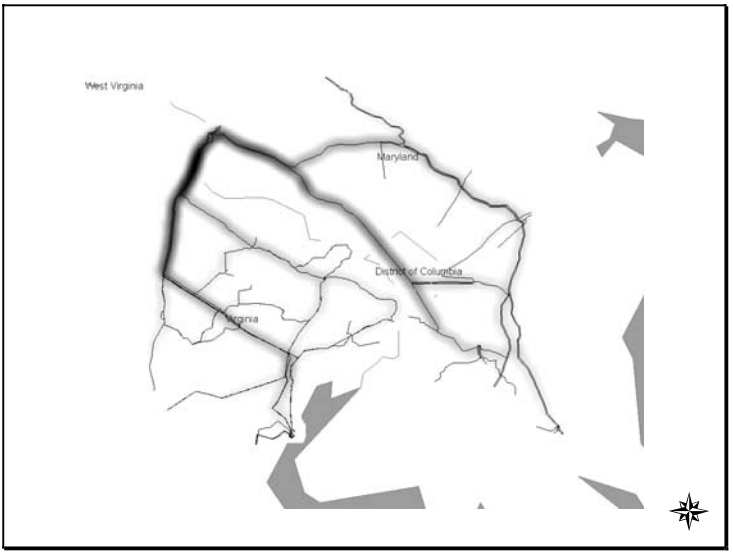


Fig. 14.1. A Structural Analysis of a Regional Infrastructure

The darkness of the shading in Figure 14.1 reflects the value of betweenness, with darker, more pronounced shading indicating areas of higher criticality and lighter shades of gray less critical. The analysis clearly illustrates the large bottleneck between the top and bottom halves of the infrastructure. A baseline can also be calculated of infrastructure's current resiliency based on the number of routes available to infrastructure. After a failure the average route length can increase, the number of available routes can decrease, and parts of the network can be disconnected. The addition of mitigation can decrease average route length, increase the number of alternate routes, and provide continuity in the case of a failure. For instance below is an example of mitigation to the infrastructure above, demonstrating how vulnerability was diffused by mitigation.

Fig. 14.2. Structural Analysis of Mitigation

The impact of a mitigation strategy was also analyzed. Figure 14.2 shows the results of this analysis. The circle identifies the additional route has been added, which now provides a method to route around the previous bottleneck. The location for the mitigation was selected by looking for existing rights-of-ways. This was done in ArcView 3.3 by overlaying a polyline theme for the Washington, D.C. area natural gas pipeline network onto the density map shown in Figure 14.2. The location selected for the mitigation is along an area where a natural gas pipeline transects the electric power gird near the bottleneck identified in the initial density map.

The structural analysis of the network with the mitigation illustrates a reduction in the criticality of the old route, but also a heavy emphasis on the criticality of the new route. This can be clearly seen when the difference in connectivity is calculated before and after the addition of the new route as seen in figure 14.3. The mitigation has effectively diffused vulnerability in the old route and provided a second route to keep all the critical assets in the previous failure connected. The white hues along the network represent reductions in betweenness scores while the heavier black shade areas are increases in betweenness scores.

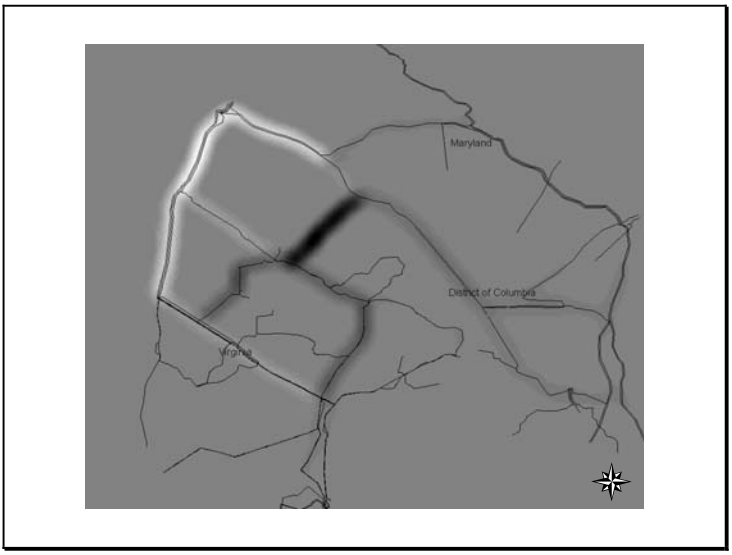


Fig. 14.3. Change in Structural Criticality After Mitigation

Table 14.1 provides some summary statistics for the betweenness scores in the network before and after the mitigation. These metrics seem to suggest that the mitigation has resulted in some improvement in the resiliency of the regional power grid. The mean and median of the scores decreased significantly implying that criticality of a few segments was diffused through the network with the addition of a new route. The variability in the betweenness scores measured by standard deviation and degree of skewness also decreased**.**

Table 14.1. Summary Statistics of Betweenness Scores Before and After Mitigation

Statistic	Before	After	Percent Change
Mean	0.1262	0.0957	-24.21%
Median	0.0526	0.0487	-7.40%
Standard Deviation	0.1714	0.1255	-26.76%
Skewness	2.1295	2.3414	9.94%

To visualize and quantify changes in the distribution of the scores, and changes in resiliency, the ranked scores are plotted for each network and an exponential model is fit to each series. Power law models were also fit however their goodness-of-fit measures were much lower than the exponential models.

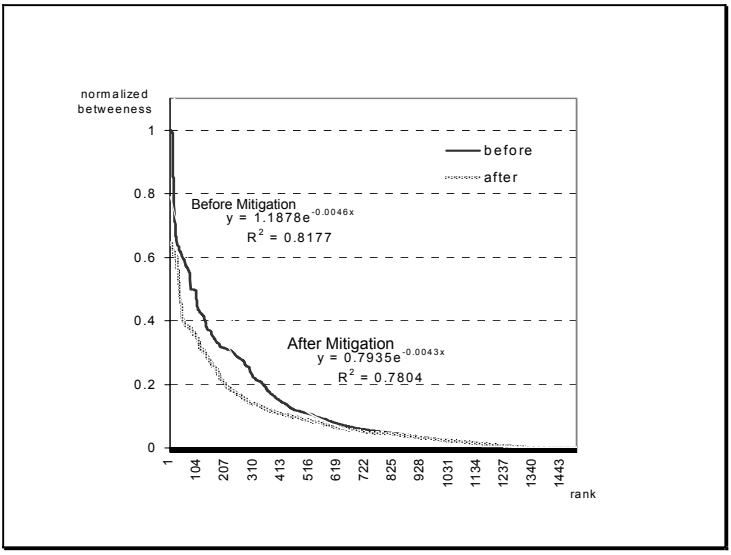


Fig. 14.4. Distribution of Betweenness Scores Before and After Mitigation

Figure 14.4 shows that with the mitigation the distribution of betweenness scores declines more gradually and this is also reflected by the decrease in the value of the coefficient on the power-law equation. According to this analysis, the addition of a new transmission line results in a 6.52% increase in the resiliency of the network. The cost of this addition can next be considered to measure the cost-effectiveness of the mitigation strategy. As a hypothetical, suppose trenching and building an additional route averages around $50,000 per kilometer. The mitigation in the scenario above was 15.6 km and would cost roughly $780,000. Using this figure, the cost-effectiveness ratio for every $1,000 spent is 0.0084%. This

baseline can then be used to compare other possible mitigation strategies to determine which provides the greatest return on investment.

Some of the advantages to the method introduced in this paper are that it is fairly easy to implement, flexible enough to be applied to different types of networks at varying geographic levels and it can be used to measure resiliency in terms of interdependencies between different network types. The latter can be accomplished by created weighted betweenness scores for each grid cell that reflect the betweenness values for each of the networks being analyzed.

One of the limitations of using the grid-based approach introduced here relates to the curse of dimensionality. As the density of a network increases, the grid resolution need to adequately capture the network topology expands and the computational power to run simulations grows exponentially. A segment-based approach, where the network topology is defined in terms of link intersections can be used as alternative to minimize dimensionality issues although this is still an issue for any relatively large network.

Further research should focus on discovering methods for reducing the dimensionality problem and also examine other metrics for measuring resiliency. This can include metrics that incorporate the operational features of the network, such as traffic levels, or impacts on the users of the network.

14.6 Conclusion

The predominant paradigm of critical infrastructure protection has struggled with obtaining buy in and cooperation from the private sector around the globe. With such a large portion of critical infrastructure owned by the private sector this is a particularly important hurdle to clear going forward. This paper has proposed a shift from protection to resiliency as one avenue to clear the difficulties faced by the current approach. The new approach, though, must go beyond simply changing language and provide an integrated methodology that allows the business case to be made to the private sector to invest in critical infrastructure. Unless a sound argument can be made that there is a financial incentive to invest then there is unlikely to be much headway made in securing critical infrastructure. To help motivate the private sector public policy tools can be used to incentivized infrastructure owners, but these too much be analyzed to determine their cost effectiveness. While there is no guarantee the new approaches will work it at least provides a baseline and metrics by which strategies, approaches, and policies can be made.

References

Albert, R., H. Jeong and A.-L. Barabás.i 2000. Attack and error tolerance in complex networks. *Nature*, **406** (6794), 378–382.

Bathelemy, M. 2003. Cross-over from scale-free to spatial networks. http://arxiv.org/abs/cond-mat/0212086.

Borgatti, S.P., M.G. Everett and L.C. Freeman. 2003. Ucinet 6 for Windows: Software for social network analysis. Harvard Analytic Technologies.

Callaway, D.S., M.E.J. Newman, S.H. Strogatz and D.J. Watts. 2000. Network robustness and fragility: percolation on random graphs. *Physical Review Letters*, **85** (25), 5468–5471.

Chen, Q., C. Hyunseok, R. Govindan, J. Sugih, S. Schenker and W. Willinger. 2001. The origin of power laws in Internet topologies revisited. *Proceedings of IEEE Infocom 2002*, **2**, 608–617.

Cohen, R., K. Erez, D. ben-Avraham and S. Havlin. 2001. Breakdown of the Internet under intentional attack. *Physical Review Letters*, **86** (16), 3682–3685.

Freeman, L.C. 1979. Centrality in social networks: conceptual clarification. *Social Networks*, **1**, 215-239.

Garrison,W. 1960. Connectivity of the interstate highway sy*stem. Papers and proceedings of the Regional Science Association*, **6**, 121-137.

Goh, K.I., B. Kahng and D. Kim. 2001. Universal behavior or load distribution in scale-free networks. *Physical Review Letters*, **87** (27), 278701-1-4.

Gorman, S.P. and E.J. Malecki. 2000. .The networks of the Internet: an analysis of provider networks. *Telecommunications Policy*, **24** (2), 113–134.

Gorman, S.P. and R. Kulkarni. 2004. Spatial small worlds: new geographic patterns for an information economy. *Environment and Planning B*, **31** (2), 273–296.

Grubesic, T.H., M.E. O'Kelly and A.T. Murray. 2003. A geographic perspective on telecommunication network survivability. *Telematics and Informatics*, **20** (1), 51–69.

Haggett, P. and R. Chorley. 1969. *Network Analysis in Geography*, New York, NY, USA: St. Martins Press.

Kansky, K. 1963. *Structure of Transportation Networks: Relationships Between Network Geometry and Regional Characteristics*, University of Chicago, Department of Geography, Research Papers.

Lakhina, A., Byers, J.W., Crovella, M.and I. Mattra. 2002. On the geographic locations of Internet resources. http://www.cs.bu.edu/techreports/pdf/2002-015-internet-geography.pdf.

Longcore, T. and P. Rees. 1996. Information Technology and Downtown Restructuring: The Case of New York City's Financial District. *Urban Geography* **17**, 354-372.

Malecki, E.J. 2002. The economic geography of the Internet's infrastructure. *Economic Geography*, **78** (4), 399–424.

Network Reliability and Interoperability Council (NRIC). 1997 .*Final Report of the Network Reliability and Interoperability Council*. Washington, DC: Federal Communications Commission.

Nyusten, J.D. and M.F. Dacey. 1968. A graph theory interpretation of nodal regions. In B. Berry and D. Marble, *Spatial Analysis*, Englewood Cliffs, NJ, USA: Prentice Hall, 407–418.

Obrien, C. 1992. Global Financial Integration: The End of Geography. Chatham House/Pinter, London.

O'Kelly, M.E. and T.H. Grubesic. 2002. Backbone topology, access, and the commercial Internet, 1997 – 2000. *Environment and Planning B*, **29** (4), 533–552.

Office of Technology Assessment. 1995. *The Technological Reshaping of Metropolitan America.* Washington, DC: US Government Printing Office.

Power, D. 2002. IT and institutions in the structuring of European finance: Urban impacts. *Economic and Industrial Democracy*. **23** (3), 335-356

Schintler, L.A., Gorman, S., Reggiani, A., Patuelli, R. and P. Nijkamp. 2005. Small-world phenomena in Communications networks: A cross-atlantic comparison. In A. Reggiani and L. A. Schintler, *Methods and Models in Transport and Telecommunications*, Berlin: Springer-Verlag, 201-219.

Taffee, E.J. and H.L. Gauthier. 1973. *Geography of Transportation*, Englewood Cliffs, NJ, USA: Prentice Hall.

Townsend, A. 2001. Network cities and the global structure of the Internet. *American Behavioral Scientist*, **44** (10), 1697–1716.

Urban, D. L. and T. H. Keitt. 2001. Landscape connectedness: a graph-theoretic perspective. *Ecology*, **82,** 1205-1218.

Warf, B. 1989. Telecommunications and the globalization of financial services. *Professional Geographer*. **41** (3), 257-271.

Wheeler, D.C. and M.E. O'Kelly. 1999. Network topology and city accessibility of the commercial Internet. *Professional Geographer*, **51** (3), 327–339.

Yook, S.H., Jeong, H., and A.L. Barabasi. 2002. Modeling the Internet's large-scale topology. *Proceedings of the National Acadamy of Sciences (PNAS)*, **99**, 13382-13386.

Index