561 LECTURE NOTES IN ECONOMICS AND MATHEMATICAL SYSTEMS

Hagen Bobzin

Principles of Network Economics

Springer

Lecture Notes in Economics and Mathematical Systems 561

Hagen Bobzin

Principles of Network Economics

Springer

Author

Dr. Hagen Bobzin
Private Docent
University of Siegen
School of Economic Disciplines
Department of Economics
57068 Siegen
Germany
hug.bobzin@t-online.de

ISSN 0075-8442
ISBN-10 3-540-27693-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-27693-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera ready by author
Cover design: *Erich Kirchner*, Heidelberg

Printed on acid-free paper 42/3130Jö 5 4 3 2 1 0

To my wife

Preface

Almost all economic activities in modern societies are scattered through space and time. Transport processes, as a consequence, pervade everyday life and they have deep impact on economic and social prosperity. Today's standard of living would just be unthinkable in the absence of water or power supply systems; road, railway, and air traffic systems are virtually used by everyone; and many people even cannot imagine to live in a world without telecommunications networks — including television, telephony, and the Internet. All examples have some kind of a transport process in common, by which people, commodities, or just data are moved along the interconnections of a network.

The purpose of this text is to provide an *economic view* on basic principles of transportation related network activities. In doing so the analysis is not restricted to certain types of transport networks at the outset and this requires a relatively simple *production technology*. In order to describe the behavior of groups of actors involved in the transport process, *microeconomic theory* suggests to distinguish between the provision of networks and of network services. Consequently, the analysis refers to at least one network carrier who offers a system of network components. On the basis of this network a second group of actors produces services and supplies them to the third group – the consumers – on the respective market. Having consumer sovereignty in mind, the principal question arises as to how to adjust the production structure of networks such that they fit best to the needs of the society.

Apart from technical problems which are caused by the complexity of network problems and by technological peculiarities, various difficulties of network analysis result from *imperfections of markets* for network services. Network carriers, for example, frequently possess remarkable market power and there are numerous externalities not only external to the network but also internal to the network. It is quite obvious that simultaneous network processes share certain network resources and overstraining these capacities induces congestion at certain places in distinct periods of time. Moreover, durable and/or indivisible investment goods cause tremendous fixed and overhead costs that must be borne not only by network participants but also by persons outside of the network.

Following the concept of the *market mechanism*, where market prices coordinate economic activities in order to utilize scarce resources efficiently, the analysis starts with the core problem of traffic assignment. Pricing strategies are discussed which ensure the efficient use of given network resources in different settings. In the next step network design abandons the assumption of a fixed network. Investment problems are examined in order to adjust the network to consumer needs; they are expressed, e.g., by travel time, accessibility, and reliability. Synchronization of network processes is an additional attempt to overcome problems associated with stochastic traffic flows. Such problems are mainly reflected by congestion and queuing which considerably reduce the perceived quality of service.

The author's deep gratitude goes to Professor Dr. Walter Buhr, University of Siegen (Germany), for his advising and supporting guidance. Acknowledgment is also to be paid to Professor Dr. Karl-Josef Koch, University of Siegen (Germany), for reviewing this book. Further thanks are due to PD Dr. Thomas Christiaans, whose comments led to many improvements of the work. Most of the burden, however, was born by my wife, Dr. Gudrun Bobzin. She offered constant intellectual inspiration and valuable criticism regarding my research on transportation economics.

Hennef (Sieg), Germany

Hagen Bobzin
April, 2005

Contents

List of Figures

List of Tables

List of Symbols

Bold small letters ($\mathbf{a}, \mathbf{b}, \dots, \boldsymbol{\alpha}, \boldsymbol{\beta}, \dots$) denote column vectors.
Bold capital letters ($\mathbf{A}, \mathbf{B}, \dots, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \dots$) denote matrices.
Calligraphic capital letters ($\mathcal{B}, \mathcal{D}, \mathcal{E}, \dots$) denote sets.

$<$	$\mathbf{x} < \mathbf{y} :\Longleftrightarrow$ $x_j < y_j \ (j = 1, \dots, n)$
$\leqq$	$\mathbf{x} \leqq \mathbf{y} :\Longleftrightarrow$ $x_j \leqq y_j \ (j = 1, \dots, n)$
$\leq$	$\mathbf{x} \leq \mathbf{y} :\Longleftrightarrow$ $[\mathbf{x} \leqq \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}]$
$\hat{}$	denotes optimal solutions
$\tilde{}$	refers to routes
$\oplus, \otimes, \oslash, \obslash$	operators in max-plus algebra; see p. 314
$*$	denotes conjugate convex functions, see p. 203
$\star$	denotes conjugate concave functions, see p. 203
$\emptyset$	empty set
$:\Longleftrightarrow$	equivalent by definition
cone	normal cone
Dom	effective domain of a proper function
n-Dom	effective domain of an n-proper function
E	expected value or mean
int	interior of a set
$\mathbb{N}$	natural numbers
$\mathbb{N}_0$	natural numbers and zero

Prob	probability
$\mathbb{R}$	real numbers
$\mathbb{R}_\oplus$	real numbers augmented by $-\infty$
rint	relative interior of a set
$\mathbf{1}$	vector with all components equal to one
∂f	subdifferential of f
∇f	gradient of f
$\nabla \mathbf{f}$	transposed Jacobian matrix $\mathbf{J_f}$
$\lVert \mathbf{x} \rVert$	Euclidean norm
$\alpha(e), \ \alpha(\boldsymbol{\rho})$	start node of arc e or path $\boldsymbol{\rho}$
$\beta_0, \ \beta_1$	parameter
$\gamma(e)$	arc weight per unit of flow $\varphi(e)$
$\gamma_e, \ \boldsymbol{\gamma}$	link prices
$\tilde{\gamma}_\rho, \ \tilde{\boldsymbol{\gamma}}$	route prices
$\delta(\cdot, \mathcal{C})$	indicator function of set $\mathcal{C}$
$\delta_e, \ \boldsymbol{\delta}$	capacity enlargements
$\boldsymbol{\Delta}$	arc/path incidence matrix defined on p. 60
Δf	superdifferential of f
ε	zero element in max-plus algebra

ζ, $\boldsymbol{\zeta}$	circuit, cycle
θ	dispersion parameter
θ_e, $\boldsymbol{\theta}$	fractions of flows using link e
$\tilde{\theta}_\rho$	fractions of flows using route ρ
κ_e, κ_{ij}, $\boldsymbol{\kappa}$	capacity of arc $e = (v_i, v_j)$
κ_v, $\kappa(\mathcal{S}, \mathcal{T})$	capacity of node v or cut $(\mathcal{S}, \mathcal{T})$
$\boldsymbol{\Lambda}$	origin destination/path incidence matrix defined on p. 60
$\boldsymbol{\lambda}$	Lagrange multipliers
λ	(1) arrival rate (2) eigenvalue in max-plus algebra
$\boldsymbol{\mu}$	Lagrange multipliers
μ	service rate
$\mu(\boldsymbol{\rho})$	mean weight of path $\boldsymbol{\rho}$
$\mu_{\max}(\mathbf{A})$	maximum cycle mean
$\xi_{(s,t)}$, $\boldsymbol{\xi}$	demand for trips
$\pi_{(s,t)}$, $\boldsymbol{\pi}$	shadow prices (assigned to OD-pairs)
π_k	probability $P\{K = k\}$
π	profit
ρ, $\boldsymbol{\rho}$	chain, path, route
$\rho(\mathbf{A})$	spectral radius of $\mathbf{A}$
ϱ	occupation rate
$\vert\boldsymbol{\rho}\vert_\ell$	length of path $\boldsymbol{\rho}$, see p. 312
$\vert\boldsymbol{\rho}\vert_w$	weight of path $\boldsymbol{\rho}$, see p. 312
σ_{ij}, Σ	delay at node v_j that does not reach v_i
τ_e, τ_{ij}, $\boldsymbol{\tau}$, $\mathbf{T}$	travel time or distance on link e measured in seconds
$\tilde{\tau}_\rho$, $\tilde{\boldsymbol{\tau}}$	travel time or distance on route ρ measured in seconds
φ_e, φ_{ij}, $\varphi_{ij}^{(t)}$, $\boldsymbol{\varphi}$, $\boldsymbol{\Phi}$	flow on arc $e = (v_i, v_j)$ (with destination v_t)
$\varphi^\#$	total number of flow units

Φ	set of feasible link flows Φ^D defined on p. 61 Φ^S defined on p. 61 Φ^S_χ defined on p. 191 Φ' defined on p. 179
φ_v^-, $\boldsymbol{\varphi}^-$	inflow of node v
φ_v^+, $\boldsymbol{\varphi}^+$	outflow of node v
χ	entropy (χ like chaos)
$\boldsymbol{\psi}$	0-1 variables
$\omega(e)$, $\omega(\boldsymbol{\rho})$	end node of arc e or path $\boldsymbol{\rho}$
Ω	feasible sets
$\mathbf{A}$, a_{ve}	adjacency matrix or arc/node incidence matrix
$\mathbf{B}$	arc/node incidence matrix
$\mathcal{B}$	set of OD-pairs
$\mathfrak{b}$	dual variable (assigned to the budget $\bar{I}$)
C	cost function, context specific
c_e, $\mathbf{c}$	cost per unit on link e
c_e^f, $\mathbf{c}^f$	fixed cost per unit on link e
$\tilde{c}_\rho$, $\tilde{\mathbf{c}}$	cost per unit on route ρ
$\mathfrak{c}_e$, $\mathfrak{c}$	marginal cost per unit on link e
$\tilde{\mathfrak{c}}_\rho$, $\tilde{\mathfrak{c}}$	marginal cost per unit on route ρ
$\mathcal{D}$	set of destinations
$\mathbf{D}$	destination/path incidence matrix
$\mathbf{d}$	timetable
$\mathbf{d}_j$, $\mathbf{d}$	deadline at node v_j
e	expenditure function
e, $\mathcal{E}$	edge, arc or link, set of edges ($e = 1,\ldots, m$); $e = (v_i, v_j)$
$\mathbf{e}_j$	unit vector having all components equal to zero with the exception that the j^{th} component is one; the $n \times n$ matrix $(\mathbf{e}_1, \mathbf{e}_2,\ldots, \mathbf{e}_n)$ is called the identity matrix; $\mathbf{e}_j^{\mathsf{T}}\mathbf{x} = x_j$

$f,\ \tilde{f}$	total travel cost
$\mathcal{G},\ \mathbf{G}$	graph or directed graph (= digraph)
$\mathcal{G}(\mathbf{A})$	precedence graph of **A**, see p. 311
$\mathcal{G}^c(\mathbf{A})$	critical graph of **A**
$g,\ \tilde{g} \equiv g\boldsymbol{\Delta}$	a line integral representing a cost term
$g^*,\ (g\boldsymbol{\Delta})^*$	conjugate convex of g and $g\boldsymbol{\Delta}$
$h_\rho,\ \mathbf{h}$	path flow on route ρ
H	set of feasible route flows H^D defined on p. 61 H^S defined on p. 61 H^S_X defined on p. 191 H' defined on p. 183
$\mathbf{H}_f$	Hessean matrix of f
$I,\ \bar{I}$	investment cost, budget
$\mathbf{I}$	unit matrix or identity matrix, see unit vector $\mathbf{e}_j$ redefined in max-plus algebra
$\mathcal{I}$	set of intermediate nodes
$\mathbf{J}_\mathbf{f}$	Jacobian matrix of **f**
$k,\ k\boldsymbol{\Lambda}$	a line integral representing the willingness to pay
$k^\star,\ (k\boldsymbol{\Lambda})^\star$	conjugate concave of k and $k\boldsymbol{\Lambda}$
$\mathrm{k}_e,\ \mathbf{k}$	excess capacities of arcs
$\mathfrak{k}_j(t), \mathfrak{k}(t)$	number of events at node v_j up to time t
$\mathcal{L}$	Lagrangean functions
$\mathcal{N}$	flow-network
$\mathcal{O}$	set of origins
$\mathbf{O}$	origin/path incidence matrix
$p_{(s,t)},\ \mathbf{p}$	minimum cost for travel from v_s to v_t
$p^D_v,\ \mathbf{p}^D$	consumer prices at node v
$p^S_v,\ \mathbf{p}^S$	producer prices at node v
$\mathrm{p}_{(s,t)},\ \mathbf{p}$	OD-specific shadow prices
P	production possibility set

P_Ω	projection operator defined on p. 135
$q_e,\ \mathbf{q}$	price vector of capacity units
$\mathrm{q}_e,\ \mathbf{q}$	shadow prices of link capacities
$\tilde{\mathrm{q}}_\rho,\ \tilde{\mathbf{q}}$	route specific shadow prices
$\mathcal{R}(s, t)$	set of elementary routes from v_s to v_t
$\mathcal{R}$	set of all elementary routes in the network
$\mathcal{S}$	subset of nodes including the source
$s_e,\ s_{ij},\ \mathbf{S}$	distance measured in meters
$\mathcal{T}$	subset of nodes including the sink
t_U	distance or transformation function
t_e	free flow travel time on link e
$\mathbf{t}_j(k), \mathbf{t}(k)$	date of the k^{th} event at node v_j
$\mathrm{t}_\rho,\ \mathbf{t}$	link specific tolls
$\tilde{\mathrm{t}}_\rho,\ \tilde{\mathbf{t}}$	route specific tolls
$\mathrm{u}_e, \mathbf{u}$	dual variables (assigned to links)
U	utility function
$v,\ \mathcal{V}$	node, set of nodes ($v = 1,\dots, n$)
$\mathrm{v}_v,\ \mathbf{v}$	dual variables (assigned to nodes)
V	indirect utility function
$w_e,\ \mathbf{w}$	link costs
$\tilde{w}_e,\ \tilde{\mathbf{w}}$	route costs
$w_e,\ w_{ij},\ \mathbf{W}$	weight of arc $e = (v_i, v_j)$
$\mathbf{W}^\diamond, \mathbf{W}^+, \underline{\mathbf{W}}$	auxiliary matrizes derived from **W**
$\mathrm{w}_v,\ \mathbf{w}$	dual variables (assigned to nodes)
$\mathbf{x}$	trip matrix in vector form

$x^D_{(s,t)}$, $\mathbf{x}^D$	demand for trips from v_s to v_t, trip table in vector form
$\mathbf{x}^S$	generated traffic in vector form
y^S_v, $\mathbf{y}^D$	nodal demand
y^S_v, $\mathbf{y}^D$	nodal supply
$\mathbf{z}$	eigenvector in max-plus algebra

1

An Outline of Network Economics

Networks are the veins between all communicating activities scattered through space and time. The wide range of network types includes power supply systems, telecommunications networks and road systems to give only a few instances. All examples have some kind of a transport process in common. Due to the specific importance of each type of network, it takes no wonder that there is a very extensive literature on the characterization, representation, and optimization of different networks. Most texts, however, draw the attention to a single type of network and then solve particular problems. For instance, the pioneering study of Beckmann et al. (1956) concentrates on highway traffic and railroad transportation. In comparison Bertsekas, Gallager (1992) discuss data network related problems. Other texts such as Bertsekas (1998) follow a more general approach with respect to network types, but consider solely network optimization problems such as shortest path, max-flow, assignment, vehicle routing, multicommodity flow, etc. These approaches are supplemented by works such as Jungnickel (1994), who studies the complexity of network problems and gives some algorithms for their solution. Last but not least even particular technical problems need deep insight because of their far reaching effects. One example is Kleinrock's (1975) work on queuing theory.

In consideration of the recent literature on *network economics* it seems to be important to delimit the above given analysis of *transport systems* in the broadest sense from another concept which is nowadays referred to as *economics of networks*; see Erber, Hagemann (2002). This concept addresses organizational phenomena including the design of economic institutions, where the nodes of a network are substituted by interacting economic agents. In doing so economic networks distinguish between the social neighborhoods of actors and the global community, and the corresponding literature analyzes the de- and re-linking of these social structures in the context of an emerging global society; cf. Erber, Hagemann (2002, pp. 236–237). At least in principle, these network types can be subsumed under the present analysis of transport networks by considering the transportation of interpersonal messages. Indeed, the importance of neighborhoods has been analyzed before in the context of so-called gravity models where, for instance, the attractiveness of retail facilities to their customers is described by some distance decay relationship. A variant of such *real* economic networks is defined by Economides[1] as *virtual* network which denotes a collection of compatible goods such as VHS video players or computer software that share a common technical platform. In this sense economic networks address issues of compatibility, coordination through technical standards, interconnection, e.g., the connection of formerly separated networks, and interoperability[2] of agents' decisions as well as their effects on pricing and quality of services; see Economides (1996). All these aspects pertain also to transport networks so that we call into question that new terms such as "economic" or "virtual networks" are too useful. Moreover, with regard to large-scale physical networks all of the aforementioned issues target essentially positive network externalities on the demand side, that is network users profit from increasing networks due to a rising number of users. This observation is also valid for transport networks, however, the effect is to be contrasted with scale effects on the supply side, which will be in the center of interest in our analysis. Clearly, many transport systems exhibit economies of scale, but more important is the determination (and removal) of bottlenecks that cause remarkable congestion costs as can be experienced in the daily rush hours. The economic problem is then to design reliable transport networks which are able to cope with a fluctuating and growing demand for traffic.

The purpose of this book is to provide an economic view on transportation related network activities. As the analysis is not restricted to certain types of transport networks at the outset, the first task is to work out similarities and differences between several network types. Afterwards networks will be interpreted as firms producing multiple services which are supplied on the respective market. In doing so the network problems such as flow maximization, traffic assignment or network design are restated in terms of microeconomic theory. For instance, a network provider may pursue some optimization task while he is restricted to the production technology of his network. Similarly, the network users or, more precisely, the consumers of network services will be assumed to behave in a utility maximizing way. The difficulties of network analysis result from various additional aspects. Simultaneous processes represent composite goods that have to share certain network resources. Overstraining these capacities induces congestion at certain places in distinct periods of time. Moreover, durable and/or indivisible investment goods cause tremendous fixed and overhead costs that must be borne not only by network participants but also by persons outside of the network. Although frequently denied, such economic observations are also valid for the Internet.

The structure of this book results from the following reasoning. Starting with different network types, we can distinguish at least three groups of actors involved. The first group consists of at least one network carrier who provides a system of network components, say nodes and links. On the basis of these networks the second

[1] Economides maintains an Internet site including a dictionary of terms in network economics at `http://www.stern.nyu.edu/networks/dictionary.html`.

[2] Technical standards, interconnection and interoperability are also major problems of transport systems as can be seen from container movements that use more than one transport mode.

group produces services and offers them to the third group, namely, the consumers. Unfortunately, even simple examples show that these three groups of actors cannot always be separated accurately and that further partitions are possible. The following table refers to different transport modes, that is road, railway, and aviation, where the cited groups overlap more or less. Car drivers, for example, produce services by making use of privately owned cars and consume these services on their own. In contrast railway travelers consume services offered by some railway company which is usually the owner of the railway network. Regarding aviation it is relatively simple to tell apart the three groups of actors, but aviation networks usually dissect into airports (referring to nodes) and air traffic control (referring to links). Similar distinctions apply to other networks (gas, water, electricity, telecommunications, etc.) where additional features must be taken into consideration as will be explained in Chapter 2.

Table 1.1 Introductory examples for different groups of network participants

network type	network carrier(s)	service provider(s)	consumers
	(nodes, links)	(additional devices)	
road	private or public supply	travelers with private car ownership	
		taxi drivers/public transport	passengers
		hauliers	shippers
railway	private or public railway companies		passengers
	railway network	railway services	passengers
	(tracks, stations, power)	(trains)	
aviation	airports/air traffic control	airlines (airplanes)	passengers

Having introduced the descriptive and technical principles of network analysis – i.e., the production technology – in Chapter 2, Table 1.1 suggests a procedure which stresses at first sight the supply side of the provision of network services. Road networks, however, indicate that consumer choice cannot be neglected in the analysis of network activity patterns, particularly, when travelers decide on route choice. Section 2.3.5 considers further decisions of network users such as mode or trip choice. Moreover, it suggests a procedure that decomposes the consumer behavior into a sequential decision process. This decomposition ends up with the core problem of traffic assignment, which is to be generalized up to locational choice.

According to microeconomic theory of the firm, Chapter 3 starts with the economic analysis of networks in terms of production theory. By assumption we begin in Section 3.1 with the simplest case where the network provides one homogeneous service. The first task is to find feasible *network activities* which are technically efficient in the sense that they indicate the maximum output on a fixed network with given capacities (maximum flow problem). The next step aims at the determination of a cost minimal network activity that supports a given output (minimum cost problem). Here only transport costs are taken into consideration. Given optimal network activities with respect to one homogeneous output, we then extend the analysis with respect to *multicommodity flows* where our attention will be drawn to a cost minimal *traffic assignment* as core problem; see Section 3.2. For the sake of simplicity the analysis starts with constant transport costs per flow unit and the demand for traffic is supposed to be fixed. However, taking *congestion* into account, average travel costs increase with network utilization. In this case users' route choice tends to be suboptimal because travelers ignore their disturbing effects on other network users. These external costs can be corrected in principle by price incentives, say Pigou taxes, such that the new traffic assignment results in an overall system optimum.

The traffic assignment problem will be generalized in Chapter 4 by relaxing several assumptions. In the first instance Section 4.1 abandons the deterministic traffic assignment by introducing *stochastic elements of route choice*. In this setting we are interested in the most likely traffic assignment, which can be found by the concept of entropy maximization. By analogy to microeconomic theory of the household, the next step is to give up a fixed demand for traffic; see Section 4.2. A *price sensitive demand* for trips implies that all effects of *traffic diversion* are now superimposed by effects of *traffic creation* or *traffic diminution*. In other words, any measure that improves traffic conditions causes two modifications of the prevailing traffic pattern: it redirects some parts of the existing flows and it induces additional flows. Other generalizations such as a dynamic traffic assignment or a multiclass-user assignment are indicated in Section 4.3.

Chapter 5 continues with the cost minimal traffic assignment of Chapter 3 by giving up the assumption of a fixed network structure with given capacities. With regard to microeconomic theory this is the case where a firm is allowed to adjust all of its factors of production in the long run rather than having fixed amounts of inputs in the short run. In this sense *network design* seeks optimal *investment programs* that improve traffic conditions. This analysis is done for both a fixed and a price sensitive demand for traffic. Similar as before we start with one homogeneous flow and examine the efficient use of investment resources. This includes the maximum attainable flow given the investment costs as well as the minimum investment costs to obtain a prespecified flow. Afterwards the network design problem is discussed for multicommodity flows; see Section 5.2. In what follows selected examples point out that special network features result in specific network structures. One example results from link specific economies of scale so that network operators can take advantage of bundling flows. Moreover, a stochastic demand with queuing requires modified investment strategies which result in surplus capacities in order to avoid frictions of a standstill. Similar results are discussed with respect to susceptible network components. Excess capacities are then used to ensure a minimum operability of a network even when the failure of a network component suspends certain flows.

Besides queuing network operators may also use *schedules* to coordinate interdependent network activities on the basis of precedence constraints. In this sense Chapter 6 goes back to the assumptions of a fixed network with a given demand for traffic. The analysis aims at the determination of *feasible timetables* with a tradeoff between two objectives. On the one hand, timetables should be able to cope with a certain amount of delays by virtue of buffer times and, on the other hand, they should minimize costs of waiting which are borne by the passengers. This last chapter is accompanied by the example of a railway system, but synchronization of network processes by timetables has a wide field of further applications; one particular example is the timing of road traffic lights.

Summarizing the work at hand is focusing on an economic view of transport networks. It delivers a consistent economic analysis of the decisions of network participants – users as well as providers – taking into consideration the costs of network design, service provision, and network usage.

2

Fundamentals of Networks

2.1 The Importance of Networks to the Economy

2.1.1 Markets for Network Services

(a) Actors Involved in the Network

The classification of transport networks starts with the market on which the respective services are traded. As will be discussed in Section 2.1.2, all these networks show their own peculiarities with regard to the used technology. Moreover, there are far reaching consequences that result from a proper network design, see Section 2.1.3. In the end economist want to determine an adequate pricing of network services. A couple of difficulties will be explained in Section 2.1.4.

Regarding markets of network services, Table 1.1 suggests to distinguish at least three groups of network participants that act on the demand and/or supply side of the respective market. It is useful to begin with a characterization of the two market sides, the market structure, and the notion of market equilibria. Afterwards, paragraph b explains what group of actors makes what network related decisions. Depending on the point of view the resulting activity pattern may be represented on different aggregation levels as can be found in paragraph c.

Demand. Transportation is a derived demand resulting from the desire to realize spatially separated activities. The most familiar case is to consume some commodity or service at a location that differs from where the good or service is produced. Shipping a freight from one place to another is the basic service produced on a transport network. In this sense (road) infrastructure may be seen as an intermediate input for other sectors (shippers). Depending on the type of freight, network outputs differ extensively. They range from physical commodities over electric currents to telecasts. Aside from freight movements, some economic activities take place at certain locations so that passengers have to make trips. Some standard examples are journeys to work, to shopping centers or for recreational purposes. In other cases networks such as telecommunications (or data) networks just connect points of mutual interaction. A telephone call may then be interpreted as a substitute for more expensive passenger or freight (letter) movements.

In the wide field of networks most systems produce various composite services which cannot be substituted, i.e., a connection between two points is different from connecting another pair of locations. The demand for distinct services such as telephone calls is reflected by so-called multicommodity networks. Furthermore, in many real world examples there are separated networks which are interconnected so that parts of a transportation service are performed by distinct suppliers. As an example one can think of journeys which use a combination of private and public transport modes. Another case is that of telecommunications networks where roaming denotes the instance of connecting two parties which belong to different networks. In many cases, such as power supply systems, it is even possible that network services are supplied by making use of networks which belong to immediate competitors.

Supply. One important distinction on the supply side of networks refers to *network provision* and *service provision*[1]. In the ideal case the two sides can be separated so that the network is operated by the first party, say carrier, while "hauliers" supply services on that network. For many examples such as railroad systems or power supply systems it seems to be easy to "unbundle" the production of services from operating a network or parts of it. Indeed, this division is reality for air traffic where airports, airlines and air traffic control operate in vertically separated companies. The telecommunications sector, however, is characterized by several companies offering essentially the same services. In doing so they make use of their own network as well as other networks for instance by leasing lines. Operating a network *and* supplying services is referred to as *bundling* that is the tying of one service or product to the supply of others. Leasing lines or parts of a network indicates the other case: service providers who do not have their own network are sometimes referred to as switchless resellers. They just provide services on the platforms of other network providers.

Another extreme case is the Internet, a network of (data) networks. Here many subnetworks, each of which belonging to another provider, are interconnected to a common network on which services can be supplied all over the world. The Internet has a hierarchical structure similar to road systems consisting of streets, (country) roads, and highways. Aside from combining networks in the sense of complements, networks can also be interpreted as substitutes. While, in general, flights require a journey to the airport, most train journeys can also be made by private cars.

Apart from all of the aforementioned details any supplier has to make up his mind what and how much to produce. While every service provider has to decide on the quality of service delivered to his customers, the network provider must

[1] Service providers supply network services to third parties whether on their own network or otherwise. Internet service providers (ISPs) in particular offer access to the Internet either to end-users or on a wholesale basis to other service providers. A virtual network provider possesses no network but uses the platforms (hard- and software) of other network providers to offer services.

dimension and design his network. Besides the technical aspect of running the network efficiently two other important problems are to be solved. The network carrier wants to forecast the future demand for services and the corresponding capacities needed and he is interested in estimates on the capacities (or potential supply) of his competitors. The answers are particularly decisive for potential entrants to the market.

Market Structure. Following Sharkey (1987), economists of the 19th century and early 20th century spoke of natural monopoly conditions arising both from superior efficiency of single-firm production and the undesirable consequences of excessive or "destructive" competition. By the middle of the 20th century it was recognized that many network industries such as railroads or telecommunications possess to some degree the characteristics of a *natural monopoly*. It is then also "natural" to assert that these monopolies (or oligopolies) have to some degree market power, that is they are able to raise prices above the competitive level in that market for a non-transitory period. In doing so private decision-making takes inadequate account of the "public interest" which is the justification for governments all over the world to establish regulatory agencies. In many cases the governments even believe to be the only authority that can provide networks as well as network services efficiently. Some familiar examples of publicly owned networks are road systems[2], airlines, telecommunications and postal[3] networks. Since 1980 many of such networks have been privatized where the most frequently cited objectives of privatization are (1) the improvement of the economic performance and (2) the generation of public budget revenues through sale receipts; see Hanke (1987) for more details. Again regulatory authorities[4] were needed to prevent the incumbent monopolist from abusing its market power. The most important objective is now to assure that incumbents cannot illegally prevent potential rivals from market entrance so that a situation arises as if competition prevails.

Nevertheless, in many network industries overhead costs are a significant fraction of total costs and competition in such an industry is still far from being perfect. Barriers to entry indicate an additional cost which must be borne by entrants but not by firms already in the industry; or other factors, which enable an incumbent to maintain prices above the competitive level without inducing entry. Some typical strategies of the incumbents to deter rivals from market entrance include the following list.

- Prices which do deter entry by rivals with the same technology are known as *sustainable prices*. A market in which there are no barriers to entry is known as a *contestable market*.[5] If the market is contestable, the threat of potential entry

[2] On the pros and cons of a commercial provision of roads in a competitive framework see Roth (1996).

[3] Even the constitution of the United States of America includes in the first article that the Congress shall have power to establish post offices and post roads.

[4] For further details on regulating specific industries such as electricity supply, telecommunications or railway transport in Germany see Eisenkopf (2003).

[5] On the concept of contestable markets see Willig (1987).

will force the unregulated monopolist to choose from the set of sustainable prices provided that the monopolist behaves in the sense of sustainability.

- In the beginning entering competitors operate relatively small networks so that they are frequently forced to complete services by making use of other networks which belong to their rivals. The access to remote networks – say by interconnection[6] – and customers can easily be hindered by *excessive transit* or *termination fees*. The German cartel agency (Bundeskartellamt), for example, taunts the electricity supplier RWE Net AG in a warning letter of 13. August 2002 to burden other suppliers with excessive metering and billing prices, so that new suppliers in particular will be significantly hindered.
  The same reasoning holds true for a single transit which is defined as an interconnection service that involves the use of one switch but no third party. As a special case, local loop[7] unbundling was mandated by the EU in December 2000. It requires those operators designated as having significant market power to make their local networks (i.e., the telephone lines that run from a customer's premises to the local telephone exchange) available to other telecommunications companies.
- *Excess capacities* may be interpreted as a thread to potential entrants. They indicate that an increasing demand can be served by the incumbent particularly when a decline of prices occurs after the market entry of a rival. Perhaps the tremendous excess capacities in the telecommunications backbone network can be interpreted in this way. The huge capacities, however, became feasible by the use of relatively inexpensive fiber cables and are possibly just the result of miscalculating future demand.
- Regulated monopolies increasingly find themselves operating not only in regulated markets, but in competitive markets as well where they are often accused of illegally using resources from the regulated market to stifle competition. Such *cross subsidies* are often used by multiservice firms to finance costs in one market from profits made in another. This behavior includes predatory intent and attempts at transferring market power from a regulated monopoly to an unregulated market. Proving this abuse of market power is difficult because multiservice firms argue that their advantages result from economies of scope. Nevertheless, in March, 2001 the European Commission concluded its antitrust investigation into Deutsche Post AG (DPAG) with a decision, finding that the German postal operator, a beneficiary of letter monopoly, had abused its dominant position by granting fidelity rebates and engaging in predatory pricing in the market for business parcel services.

[6] Interconnection designates the physical and logical linking of telecommunications networks used by the same or a different organization in order to allow the users of one organization to communicate with users of the same or another organization, or to access services provided by another organization. Services may be provided by the parties involved or other parties who have access to the network.

[7] Local loop refers to the access network connection between a customer's premises and the local exchange. This usually takes the form of a pair of copper wires.

> In the USA regulated railroads faced increasing competition from regulated and unregulated trucking, and regulated telephone companies faced increasing competition from private networks. In effect regulators were asked in what way a regulated firm should be allowed to compete with an intermodal rival or entrant. If an *intermodal competitor* enters only the most lucrative markets, he was attacked as "cream skimming" by the regulated firms and portrayed as innovative competition by the entrant.

Another barrier to entry results from barriers to exit. For instance, sunk costs are not recoverable if the activity for which they were incurred ceases.

Given inferior market results one has to think of proper forms of regulation. Two categories are to be distinguished. (1) *Ex ante regulation* includes the separation of networks and services (power supply systems, railway systems) or the unbundling of other parts of vertically or horizontally integrated firms (unbundling local loop in telecommunications systems). Other measures seek to guarantee the access of new competitors to the networks of incumbents at "fair" prices (formerly state-owned monopolies). (2) *Ex post regulation* focuses on the correction of market results. For example, some measures try to enforce marginal cost pricing, others impose price caps, and again others regulate the rate of return. In face of a long list of dissatisfactory and failing regulatory programs it is frequently postulated, not only by lobbies, to deregulate industries. By most economic measures, the industry after *deregulation* has been operating more efficiently and the average firm has increased profitability, see Breyer, MacAvoy (1987). Network users in particular profit from deregulation. For example, deregulating the German telecommunications sector has shown a drastic decline in prices and, correspondingly, an increasing demand; cf. Götz (2001).

Market Equilibrium. Having described aspects of both the demand and the supply side as well as the market structure, in the next step we are interested in the characterization of market equilibria. Note that the demand for network services refers to composite goods such as a trip from A to B involving the utilization of a sequence of roads. Each network resource in turn, that is for example one road, represents the supply side of the network, but it provides merely a fractional part of a complete network service. In accordance with Beckmann et al. (1956), the connecting link is found in the distribution of services over the network.

Given a meaningful definition of a network equilibrium, economists are interested in several of its properties. First, the *existence* of a technically feasible network activity is needed such that the output matches the demand under the prevailing circumstances. This situation is called a (static) equilibrium if none of the participants in the network has an incentive to alter its behavior. That is, all actors behave in accordance with correct expectations on the essential properties of the prevailing equilibrium. Second, in some special cases one can show that only one network equilibrium exists. This result on *uniqueness*, however, must be handled with care as there may be alternative representations of the equilibrium which are not unique. Third, the equilibrium is said to be *stable* if any but not too hard perturbation leads back to the equilibrium state. Provided the adjustment process can be modeled correctly, stable equilibria are particularly useful for computational purposes. Fourth, *efficiency* draws the attention to the effectiveness of resource utilization at the equilibrium. Many networks are characterized by the fact that the individual behavior incurs external effects. Congestion is a typical example for inefficiency. Fifth, when inefficiency grows to an unacceptable extent, one is interested in possibilities such as road pricing to alter the results. Changing the basic conditions of a network induces a series of reactions of the participants. For example, the supply side must take into account regulatory measures as well as the reactions of the network users.

Apart from short term equilibria, long term effects of any intervention are one of the major problems in network design. Network improvements affect not only the demand patterns for network services but can change the whole market situation. For example, an additional road, bridge or airport can shift location decisions of households and firms due to new accessibility conditions. The demand thus can be expected to increase as a whole but it grows asymmetrically. These effects are less emphasized in communication and distribution systems.

(b) Decisions

In accordance with the aforementioned behavior of network participants, in more detail, network users as well as network providers have to decide on a large variety of further aspects concerning network services. As it is essential for the market result which actor can select what action, let us distinguish again between the demand side and the supply side of the market.

Demand. In accordance with standard markets, many networks require a binary choice on participation which implies to buy some further devices such as telephones, cars, radios and so on. Afterwards the client decides on the amount of service he wants to consume at the prevailing market conditions. This amount can be divisible, lumpy, discrete or even binary. Just observe electric currents, durations of telephone calls, number of trips or radio access. Furthermore, depending on the market structure the client has to choose a service provider. In power supply systems the customers have to preselect one supplier. In telecommunications the selection of a service provider can be done call by call. Of course, for monopolies such as the German letter monopoly there is no choice left to the customers.

All of the above aspects can be detected in many markets. Some specific decisions regarding transport networks include (1) *trip choice*, (2) *mode choice*, and (3) *route choice*. The first case includes the binary choice whether to make a trip or not, as well as selecting the frequency of trips. Moreover, the traveler has to decide on the destination and the departure or arrival time of each trip. The second case refers to alternative modes of transportation such as using a private car, a public bus or the subway. Routing is a special question of the type of network at hand. In some cases the traffic pattern can be regulated by some central authority. For a broad class of transportation networks, however, the travel patterns are set up by individual users. Each traveler chooses the cheapest way to arrive at his destination irrespective of the implications on some aggregate system optimum. The wide range of solving the routing problem can be illustrated by the following list:

- Routing in railway systems with relatively low connectivity is essentially a question of the network topology. The railroad carrier then coordinates the demand for trips by suggesting synchronized timetables.
- With regard to road traffic car drivers make their own route choices. This process is more flexible particularly because there is less need to synchronize individual decisions than for railways. This flexibility is contrasted with the disadvantage that drivers tend to ignore the external effects they impose on the rest of the network. Hence, the resulting user optimal traffic assignment is suboptimal compared to an overall system optimum. This gives rise for thinking about a central authority which corrects the individual behavior by some traffic guidance system.
- Airline passenger carriers and parcel delivery networks take care of routing by developing so-called hub-and-spoke networks. At first, all flows are collected via spokes at specialized nodes, say hubs. In doing so the bundling of flows on the interhub links makes more efficient transportation technologies available to the carrier.
- For telecommunications systems, routing is a question to be answered by the network provider. Possible solutions depend on the way of transmitting data. The regular voice telephone network uses circuit-switched traffic, that is all (or fixed parts of the) resources on the communication circuit are unavailable for other parties. By contrast, most data networks are based on packet-switched traffic where small units of data are routed through a network based on the destination address contained within each packet. When packets arrive at a switching node they are stored and then forwarded at the full transmission rate as soon as the communication link is free. This method provides a more efficient utilization of network resources but requires more capabilities of the hardware.
- Due to physical conditions, power supply systems operate without routes. This is an important technical problem because the voltage over the entire network including all power stations and all power points must be constant.

Supply. Besides routing network providers have to decide on the *design* of their networks. Planning the network structure must take into account a list of further aspects. (1) The formation of nodes and their connections or in short the *network topology* determines the available variety of services. In terms of geographic coverage, for instance, a network extension may have the task to offer network access to additional customers. Similarly, it determines the potential interconnection with other (competitive) networks. (2) The capacity of a network indicates the maximum feasible throughput. This measure is hard to define for multicommodity networks. On the one hand it depends on the capacities of the individual network elements as well as their combination. On the other hand, any bundle of services has its own structure and, therefore, leads to different bottlenecks indicating different capacity measurements. (3a) The design features of a network determine also the quality of service (QoS), a collective measure of the level of service delivered to the customer. Although this term is closely related to the Internet, it has corresponding significance to any other network. QoS can be characterized by several basic performance criteria, including availability (low downtime), error performance, response time and throughput, lost calls or transmissions due to network congestion, connection set-up time, and speed of fault detection and correction. Service providers may guarantee a particular level of QoS to their subscribers. (3b) Apart from the scale of a network and QoS, accessibility[8] is another attribute determining the attractiveness of a network. Redundant connections, for instance, do not only improve the reliability but also can make the network easier to use by increasing the number of alternative routes and dispersing the traffic flows.

Interaction. The behavior of both market sides results in a list of short-run and long-run effects. Altering the features of a network changes the demand for services. This effect of traffic generation or diminution is overlapped by problems of traffic diversion. In accordance with the new network conditions, the route choice (reassignment) and/or the choice of destination (redistribution) will be modified. Hence, it is important at least for the network designer to take the response of the network users into account. The estimation of such short-run reactions is already a difficult task but it is even more complicated to forecast long-term effects. Almost all location or relocation decisions (housing, productions, warehouses, hubs, etc.) are based on accessibility criteria which are determined by the prevailing network structure. Changes in these conditions can lead to new agglomeration areas with a partially excessive utilization of network resources. Furthermore, more or less accidental events such as the politically motivated location of an airport can make network planning an absurd waste of time.

(c) Traffic Representation

Describing the process of movements on a transportation network follows the purpose to be modeled. Independent of the items carried by the transportation system two major approaches in representing traffic can be distinguished.

The *microscopic view* lays emphasis on single elements, users, and processes. Each element or part of the network has its own functionality which is in most cases restricted to the technical properties of adjacent elements. For example, capacities limit the maximum flow on links and nodes. Every network participant pursues individual tasks. Drivers, for instance, are assumed to follow utility maximizing (or expenditure minimizing) rules resulting in an individually preferred free flow speed presumably disagreed by other trip-makers. Most processes consist of a conditional sequence of steps where the failure at any stage makes the whole process unusable. When such a process arrives at a device occupied by some other process it has either to wait in a queue or it dies.

By contrast, the *macroscopic view* models traffic as a continuum akin to fluid. In road traffic, for example, one can think of two strategies to increase the throughput of a road. Either all drivers increase their speed so that followers can enter the road earlier. Or the drivers move up closer according to a compressible fluid so

[8] Accessibility typically refers to the ease with which desired destinations may be reached and is frequently measured as a function of the available opportunities moderated by some measure of impedance.

that more trip-makers can use the road at the same time. The problem is that both approaches increase traffic nuisance. This effect is referred to as congestion and requires all drivers either to increase safety distance or to slow down speed. Another phenomenon is described by shock waves. They represent the boundary between traffic of two different densities and thus of different behavior.
In between the two views lies the broad area of aggregating individual processes to operational variables such as total demand for services during the busy hour. The combination of overlapping processes, however, requires detailed work on synchronization. Besides an appropriate scheduling, some of the synchronization in waiting systems can be done by queuing. At least for loss systems queuing is ruled out.

2.1.2 Network Technologies

(a) Production

The basic ingredients of a network are nodes and links each of which having their own functionality. The combination of these network elements must take into account technical restrictions and other properties such as sharing of network resources or queuing. In the latter case the start of an activity depends on the termination of several other activities though not all activities are able to wait in a queue. Hence, the way of composing network elements determines the performance of the network as a whole.

Combinations of Network Resources. At first, observe that networks and their constitutive elements can be complements as well as substitutes.

Complements The combination of consecutive links to one path emphasizes the complementary character of arcs (and nodes). When one element fails all routes including the particular element cannot be used until it has been repaired. Similar observations are true for combinations of networks. The interconnection of telecommunications networks has already been mentioned. Joining road systems, railway systems and airlines provides another example where traffic is originated and terminated in different networks possibly using another transit network.

Substitutes Parallel links or alternative routes show that network elements or combinations of them can serve as substitutes. This property is especially useful when some parts of a network are out of service. Maintenance work, for instance, needs redundant network elements to keep the network operating. Similarly, the selection of some service provider operating on his own network shows that networks as a whole may be seen as substitutes. Another example is that of train journeys which can usually be substituted by auto trips.

As it is common to all production systems, networks of nodes and links are used to transform inputs (time, energy, etc.) into network services (flows) by some production technology the details of which are treated as a black box. Nevertheless, constructing and studying simple models requires to enumerate and define the most important variables and to specify the relevant relationships between them. One of the most essential parts of the transportation systems studied refers to the production of services. Hence, it is useful to go through a list of aspects which will be covered by the model in question or left out.

Factors of Production. Transportation systems are typically characterized by large indispensable amounts of *fixed (lumpy or indivisible) inputs* which usually cannot be resold. Just think of railway tracks, roads, fiber cables, telephone switches and so on. That may be different for other fixed inputs such as airplanes, locomotives, cars or even copper cables. Parking (and maintaining) airplanes in a desert is a remarkable case that they are expected to be reused in the long-run. Most of the above examples require maintenance work[9] as a fixed input per period. Maintenance itself may be seen as a problem of low analytical difficulty, but its importance rises from the fact that maintaining some device usually means to take this element out of service for some time. This problem could be ignored with respect to vehicles which can be substituted easily in most cases (rent a car). Maintaining switches, however, can imply that complete subnetworks are disconnected from the rest of the network. All activities that depend on the device in question must then be stopped.

Variable inputs depend on the amount of outputs produced. The most frequently used examples are time and "work" needed to move one item of given size one unit of distance at a given level of service, say speed for instance. As the physical work, that is power times distance, performed by the system has many determinants, it is much easier to measure the consumption of gasoline which serves as a store of available energy. Notice if the level of service is measured by speed, then increasing the level of service leads to a proportional decline of time need, but one can expect that the fuel use and the risk of accidents increase progressively. Last but not least, human work (taxi drivers, air traffic controllers, technicians, etc.) is usually measured by the time, say man hours, needed to perform services.

Outputs of Networks. A first classification results from the fact that networks provide either a single good (power or water supply systems) or many commodities. In the latter case each point-to-point connection (road networks or telecommunications systems) is interpreted as a separate service that shares network resources. The complementarity of network elements is reflected by output measures such as vehicle kilometers traveled per hour or year, call minutes per day, or kilowatt hours. This way of measuring network outputs is of minor importance for the consumers. They want to consume distinct network services in the sense of bulky composite goods, all other services are irrelevant. In other words, the composition of homogeneous service units may lead to a large variety of differentiated goods.

In any case product differentiation is a common strategy of network and service providers to establish some degree of market power. Network providers can set technical standards to make their network incompatible with other networks, cf. different gauges of tracks. Service providers use similar strategies of horizontal

[9] Costs to maintain existing infrastructure involves (1) ordinary maintenance, e.g., cleaning and winter maintenance, which is independent of road use and (2) maintenance, which depends on the volume of vehicles, e.g., surface dressing.

product differentiation to tie customers to themselves, cf. the supply of natural gas with different calorific values. Other examples refer to travel time, risk, comfort, reliability, etc. Furthermore, vertical product differentiation is common to almost all networks. Passenger carriers as well as packet delivery companies offer different levels of service to particular groups of customers and grant rebates to major clients.

Apart from product differentiation network services include two major classes:[10]

- communication between individuals or groups where the service "merely" consists of establishing end-to-end connections (telephone services, broadcasting programs);
- movements of passengers or cargo through an area (journeys or trips by different transport modes, parcel delivery, postal services).

Most services differ with respect to time and space. The *service time* has two essential aspects; it is related to the duration of service (cf. telephone calls, travel times) and to the point of time when the service is made available (departure times). Busy hours and particularly peak loads put high pressure on the given capacities. Contrasted with time we have to analyse *space*. Many economic activities are tied to certain locations requiring trips of different length. Distance is also a problem of covering space. Most networks serve in the beginning only agglomeration areas and leave out the periphery so that rural areas have no or only limited access. In contrast, many regulated network companies are obliged by regulators to provide universal services all over the space, cf. Deutsche Post AG. As a service in turn, these companies are protected against rivals which are not willing or able to provide universal services.

In the above sense *accessibility* can be added to the aforementioned classes of network outputs. Although there are various definitions of accessibility two major categories of indicators can be distinguished, cf. van Wee et al. (2001). (1) Infrastructure related approaches focus on the characteristics of infrastructure and on the flexibility of its use. For example, the density of networks is measured by total length of a network in relation to space, and coverage indicates the fraction of space or the nuumber of potential customers having access. Flexibility refers, for instance, to choice of departure time, speed on motorways, the opportunity to avoid congestion and so on. (2) Activity related indicators draw attention to needs such as living, working, recreation or shopping reachable within a certain time limit or distance.[11] One measure may be the population density, i.e., the number of inhabitants per square kilometer. The worth of accessibility becomes clear if we think of some network provider who rents out parts of his network with accessibility as one performance criterion.

[10] Nagurney, Dong (2002c) add economic and financial networks as a third class besides transportation networks and communication networks. They discuss two examples of supply chain networks, namely, the online grocer Tesco and the book retailer Amazon.com.

[11] Gravity models are used to include time and distance properties; they combine the attractiveness of some point or areal zone with a distance decay functions in order to explain spatial interaction patterns for example between central business districts and their periphery.

Some further aspects on the description of network outputs result from the fact that network services often represent *composite goods*. In most cases network services are the result of many sequential or concurrent steps. For example, a journey may consist of several subsequent trips possibly by different transport modes. It is then plausible when analyzing parts of a network to speak of intermediate goods. A railway station, for instance, can be used to initiate or to terminate a trip. But it also serves as a point of interconnecting trains in which case arriving and departing trains are interpreted as inputs and outputs, respectively. The arriving train provides a transport service needed to complete a set of different journeys. This observation leads to another problem when calculating the contribution of one network element to the provision of one unit of a final network service.

2.1 Remark (Quality of Service, QoS) Competition in the provision of communication services has forced the providers to measure their quality of service in an operationally meaningful manner and to guarantee predefined levels of service to their customers. The following list of QoS criteria applies to (tele-)communications systems (see Hardy (2001)), but it has corresponding meanings for other networks such as road systems, where trip times, risk of accidents, and reliability are of major importance.

Latency is the delay in a transmission path or in a device within a transmission path. In a router, latency is the amount of time between when a data packet is received and when it is retransmitted.

Jitter refers to the distortion of a signal as it is propagated through the network, where the signal varies from its original reference timing and packets do not arrive at their destination in consecutive order or on a timely basis, i.e., they vary in latency. In packet-switched networks, jitter is a distortion particularly damaging to multimedia traffic.

Bandwidth indicates the theoretical maximum transmission capacity of a computer channel, communications line, or bus. As the theoretical bandwidth is approached, negative factors such as transmission delay or signal-to-noise ratio can cause deterioration in quality, which corresponds to congestion in road traffic.

Reliability addresses the need of error-free communication, where each signal produces noise in its surroundings and deranges other signals. These errors can be corrected up to a certain amount. Beyond this point messages cannot be guaranteed to be delivered without errors.

QoS differs from grade of service (GoS) which addresses accessibility and availability of network services. Accessibility refers to the possibility of conditions that make it impossible to set up end-to-end connections normally supported through a telecommunications service. While accessibility is a measure of perceived QoS, availability measures intrinsic QoS which is defined from the viewpoint of the service provider rather than the service user. Availability is a ratio of uptime to total time. Network performance signifies not only QoS such as overall system delays but also throughput, both of which depend among others on the speed on the communication lines, number of hops, and occupation rates. □

Production Technologies. Networks exhibit positive and negative consumption and production *externalities*. A positive consumption or network externality signifies the fact that the value of a unit of the network commodity increases with the (expected) number of units sold; see Katz, Shapiro (1985) or Economides (1996). The benefits of a telecommunications network rise with the number of customers having access. This positive externality of an extensive network utilization is contrasted with congestion effects when the network is used more intensively. Almost all flow units disturb the traffic in their immediate surrounding. As traffic becomes heavier increasing flows use the same resources at the same time so that the mutual disturbances cannot be evaded. When congestion predominates, the corresponding flows are blocked and the movements are stalled. Congestion is usually a temporary problem due to substantial variability in demand. During most of the operational time of the system, that is outside the relatively short periods with peak loads, one has considerable excess capacities.

Both cases, congestion and excess capacities, indicate an inefficient resource utilization although sharing of a network resource is in principle possible. However, even uniform flows can cause severe problems when concurrent processes do not match. A typical strategy to *coordinate* road traffic flows is to control junctions by traffic lights and to limit the speed on the connecting roads. In an ideal (deterministic) world this procedure avoids congestion effects and on average all drivers can realize a certain speed. Another strategy is *synchronization*. In railway systems schedules represent the conditioning of events. A departing train must wait for certain incoming trains so as to allow passengers to change. By contrast, limiting capacities bind the number of trains waiting at a station. The preceding approaches show that stopping processes and storing them in waiting rooms until they are served cannot be avoided in any case. However, we have to distinguish

loss systems where telephone calls get lost if they are blocked by occupied facilities, and

waiting systems where incoming jobs are able to wait in queues. Customers may be patient and willing to wait but others may be impatient and leave after a while. For example, in call centers, customers will hang up when they have to wait too long before an operator is available, and they possibly try again after a while.

When *queuing* as a symptom of congestion occurs, the operator has to think about strategies how to minimize the waste of time of his clients.

- Queuing systems are understood as (1) a link connecting a service station with a subsequent (2) waiting room, where the job has to wait until it is handled by the following (3) server. Accordingly the time need consists of a propagation delay, a queuing delay, and a service time. As the propagation delay usually cannot be influenced, the waiting time in the queue plus the service time are referred to as sojourn time or lifetime of the job.
- When jobs (passengers, data packets, etc.) arrive they may have to wait behind other jobs. The arrival process introduces random delay. The length of the queue grows if the number of incoming jobs exceeds the capacity of the server. As the devices have finite buffering, arriving jobs are lost when the queue buffer is full.

- Queue disciplines determine the order in which waiting jobs enter service. Two examples are first in first out (FIFO) and service in random order (SIRO). Furthermore, priorization is often used to improve efficiency. Rights of way rule road traffic. In data networks different jobs can be privileged to use certain parts of a capacity. Jobs can even be rejected to reserve capacities for other jobs of higher priority.
- The relevant *performance* (or delay) *measures* in the analysis of queuing models are the distribution of the waiting time and the sojourn time of a customer.

Regarding data networks several *congestion management tools* may be implemented to handle an overflow of arriving traffic. They sort the traffic and determine some method of prioritizing it onto an output link. Analogous concepts are used in other types of transport networks.

First In First Out (FIFO) Queuing in its simplest form involves storing packets when the network is congested and forwarding them in order of arrival when the respective network element becomes free. FIFO is the default method of queuing in many instances, but it has several shortcomings. Most importantly, FIFO queuing makes no decision about packet priority; the order of arrival determines promptness and resource allocation. Moreover, it does not provide protection against ill-behaved jobs. Bursty sources can cause high delays in delivering time-sensitive application traffic, and potentially to network control and signaling messages. FIFO queuing was a necessary first step in controlling network traffic, but today's intelligent networks need more sophisticated approaches.

Priority Queuing (PQ) ensures that important traffic gets the fastest handling at each point where it is used. It was designed to give strict priority to important traffic. PQ can flexibly prioritize according to traffic mode, incoming interface, packet size, source/destination address, and so on. In PQ each packet is placed in a separate queue based on an assigned priority. During transmission, the algorithm gives higher-priority queues absolute preferential treatment over low-priority queues. This is a simple and intuitive approach, where higher-priority traffic is usually rate limited to protect lower-priority traffic against distortion or disruption.

Custom Queuing (CQ) guarantees bandwidth. CQ was designed to allow various applications or organizations to share the network among applications with specific minimum bandwidth or latency requirements. In these environments, bandwidth must be shared proportionally between applications and users.

Weighted Fair Queuing (WFQ) is a queuing method for situations in which it is desirable to provide consistent response time to heavy and light network users without adding excessive bandwidth. It is a flow-based queuing algorithm that does two things simultaneously: it schedules interactive traffic to the front of the queue to reduce response time, and it fairly shares the remaining bandwidth between high bandwidth flows. WFQ ensures that queues do not starve for bandwidth, and that traffic gets predictable service. Low-volume traffic streams – which comprise the majority of traffic – receive preferential service, transmitting their entire offered loads in a timely fashion. High-volume traffic streams share the remaining capacity proportionally between them.

Each queuing algorithm was designed to solve a specific network traffic problem and has a particular effect on network performance.

Examples. The extremely different inputs, outputs, and production technologies can be explained best by giving some examples.

Water resource planning or *hydraulic systems* represent a relatively easy case in the sense that they distribute[12] one homogeneous and perfectly divisible good (water, gas, oil). Hence, routing has no importance; merely sources and sinks are of interest. Notice, however, that natural gas with different calorific values cannot be mixed. The nodes of such networks correspond to production platforms of limited number, pumping stations, reservoirs or lakes, which are connected by pipelines. Typically these networks have low connectivity implying that failures of single elements may require to shut down large parts of the network for a relatively long time; that may be weeks for gas networks.

At first glance *power supply systems* seem to be very similar to hydraulic systems. The decisive difference, however, is that electricity cannot be stored without very inefficient conversion. Furthermore, there are various technologies using different energy sources such as gas, fossil fuels, nuclear power, or hydroelectric power. Gas turbines in particular have low capital but high generation costs, while nuclear power stations have the converse properties. Because nuclear power systems are inflexible in the adaption of a varying demand, gas turbines are mainly used for peak loads. Note also that the use of hydroelectric systems – with low generation and capital costs – depends strongly on the multi-purpose of water and on water inflows. Power supply systems usually consist of many power stations and the network is highly meshed so that failures have on average less drastic consequences.

Television or *radio* companies broadcast a variety of programs over their networks. The transmission requires either a network of cables or follows wireless via air or satellite. In principle many customers receive the same service from one sender. Of course, the quality can vary depending on the hardware used. Radio networks are typical loss systems in the sense of queuing theory, although complete transmissions can be stored by using recorders.

Telecommunications systems are multicommodity networks as they connect a large variety of origins and destinations. In a communication network, nodes represent origin and destination stations for messages (also computers, satellites, switches, etc.), and arcs represent transmission lines (copper cables, fiber optic links, microwave links, etc.). Flows include voice messages, data, video transmissions, etc. As has been mentioned above, the regular voice telephone network uses circuit-switched traffic, but modern solutions use more and more often store-and-forward switching which admits a more efficient utilization of network resources. In any case routing and coordination of traffic are tasks of the system operator. Customers merely await that telephone calls will be established and held over a certain period of time, where the quality of transmission must be acceptable. In data networks, transmission

[12] Water supply systems are to be supplemented by sewage disposal systems.

has different further requirements on quality: completeness, transmission speed, minimum bandwidth, maximum delay, reliability, security, etc.

Railway systems are multicommodity networks in the above sense of telecommunications systems. Moreover, product differentiation leads to many levels of service: first and second class, slow and fast connections, etc. The nodes represent stations, switches or rail-yards and the arcs are tracks. (The transport network of *inland navigation* mainly consists of ports and shipping canals.) The flows correspond to passengers or cargo loaded on trains with maximum capacity. As railway systems have low connectivity there is only a limited number of meaningful routes for each origin destination pair. At the same time the concurrent and sequential processes require high efforts of synchronization. Some flexibility remains with regard to waiting systems. Even impatient passengers are able and willing to wait as long as there are no other more attractive alternatives.

The *airline industry* is very similar to a railway system. What makes them different in reality is the market structure. Most airports are under the control of special companies selling – among others services – time slots in which this network node can be used by the respective carrier or airline. Furthermore, the air traffic control coordinates the aircrafts using the same airspace at the same time to ensure safety.

Road networks are much more flexible. At least partially they can be used simultaneously by many different transport modes (pedestrians, bicycles, cars, buses, trucks, etc.) with extremely different characteristics. The nodes are given by junctions, parking places, service stations and so on. Trip-makers individually decide on modes, routes, departure times and speed including take overs possibly at the risk of oncoming traffic. Disturbances are omnipresent and a particular problem at junctions with self-coordination or traffic lights. A long list of rules exists to ensure safety (e.g., drive on the "right" side) besides operational motives.

Theoretical Implications. All networks have in common that they consist of nodes which are connected by links. These links may be undirected edges or oriented arcs. Both, nodes and edges, are usually characterized by capacities imposing upper bounds on the respective flow per period. Furthermore, weights such as travel time or length are assigned to the branches. Given a predefined level of service, the weights together with the utilization of capacities will determine the link cost per flow unit.

The main problems result from the fact that many processes are performed by the network at the same time. Some of these processes pass successively, while others occur simultaneously. (1) Splitting a process into a sequence of consecutive jobs requires that certain conditioning activities are finished; initiating a job also requires free and operating facilities; and eventually the output of a process needs either free storage capacities or access to the subsequent device. (2) Many processes such as trips overlap so that concurrent jobs may use the same resource at the same time (passengers in a train). When the device in question can handle only one job at a time, jobs have to wait in queues until they can be served one after the other (landing aircrafts). Such processes need typically some sort of cooperation (safety distance) or coordination (drivers at junctions). When the restrictions due to overlapping become stronger more effort is needed for the synchronization of processes (time tables for trains).

Modeling the production structure is the first major problem in network analysis. In view of microeconomic theory the task is to develop a construct similar to the "production function" of a firm producing a single homogeneous good. Such a production function describes technically efficient activities where the maximum output is achieved at given inputs and/or the given output is produced by minimum inputs. Although it can even be difficult to detect efficient activities for firms producing one homogeneous commodity, there are a lot of essential obstacles regarding network technologies which require a broader approach. Most networks produce a large variety of outputs by a large number of concurrent processes. A substantial part of the factors of production is indivisible or at least lumpy; take, for example, a bridge of fixed length where the capacity can be varied within limits by the number of lanes. Durable inputs require far reaching investment decisions resulting among others in huge fixed capacities which are idle except for temporary peak loads. This is true for almost all parts of networks, cf. cars, fiber cables or airports. Besides investments and a substantial variability in demand further aspects of time grow in importance. When many interacting processes make use of the same network they need synchronization in order to avoid a waste of resources. Furthermore, technical progress is hard to realize in the sense that networks with an old technology may be incompatible with recent developments. It is not unusual to observe a mixture of old and new techniques (telecommunications systems) or incompatible systems side by side (railway systems). In view of all these network attributes it takes no wonder that many efficiency problems cannot be solved particularly for large scale networks.

Throughout the book multicommodity flow problems will be in the center of interest even though networks producing a single output are discussed for explanatory purposes. Multicommodity networks generate a wide range of services where each commodity will be interpreted as connecting pairs of not necessarily adjacent nodes. For example, in road traffic a trip from one location to another leading through different roads and junctions represents one commodity and the collection of all commodities is referred to as trip table. In telecommunications systems each commodity corresponds to a so-called end-to-end connection.

As technical progress goes beyond the scope of this book, the analysis focuses on the factors of production needed to generate a bundle of services at a given technology. Three major classes of inputs will be distinguished:

fixed inputs with regard to the topology of the network, e.g., roads, bridges, airports, switches, routers, etc.;

fixed inputs referring to the provision of services, e.g., cars, trains, airplanes, etc.;

variable inputs depending on the amount of service, e.g., fuel, electric currents, time, etc.

The first two classes indicate the *investment problems* of the network and the service providers. The main problem, however, is to determine the contribution of the above inputs to the services produced. This may be relatively easy in the case of variable inputs but at least for network components it is hard to assign them to specific services. As most network components accommodate many processes and every process uses a sequence of components it is useful to think of intermediate goods as a fourth class of inputs. Then, each commodity can be interpreted as a composite good consisting of a series of services produced by the individual components of the network in question.

Decomposing networks this way allows to analyze the performance of each network element, cf. the capacity of a single road. The problem is now to go one step back and to analyze the cooperation of all elements constituting the network possibly with respect to some kind of "optimal" network utilization. One example is to detect the maximum throughput of a network when all link capacities are known. By intuition, the maximum flow is determined by the some "bottleneck" usually consisting of more than one network element. Just think of several parallel congested bridges connecting two parts of a town limiting the flow from one district to the other.

Flows on a network are analyzed similar to networks and its components. Each flow unit passes through sequences of consecutive links and nodes where many routes may be available. In doing so it disturbs other flow units in its immediate surrounding. Although services supplied by road systems are sometimes referred to as public goods[13] having the property of non-rivalry, networks show a limited capability of serving flow units which simultaneously consume the same network service. Furthermore, flow units produce "noise" in the sense of external effects. When traffic grows heavier the disturbing effects induce congestion such as heat increases the resistance of a conductor. Depending on the type of network, processes in a network can be synchronized in different ways. (1) The individual flow units coordinate the behavior on their own, cf. road traffic with individual choice of routes, safety distance, speed and so on. (2) A carrier organizes flows in order to improve the utilization of given capacities, cf. railway systems with appropriate time tables or telecommunications systems using routers for an appropriate traffic assignment.

Finally, the extent of a network can change its functionality. For example, expanding a network may lead to a more uniform utilization of indivisible resources so that excess capacities can be reduced. *Economies of scale* are also realized by hub-and-spoke networks or other hierarchical networks. The idea is to bundle flows on the interhub links so that carriers can use for instance larger aircrafts. *Other* so-called *network effects* go beyond pure technical aspects. Network utilization frequently makes only sense if there is a minimum number of processes. Two examples illustrate this network characteristic. (1) The benefits of using a network increase with the number of customers or with the extent of the network. This is the case of an increasing number of clients in a telephone network. *Metcalfe's Law*[14] suggests that the usefulness, or utility, of a network equals the square of the number of users. According to Varian (1999), consumers would generally like to be connected to as large a network as possible. This implies that if there are several different providers of networks, then it is very advantageous to consumers if they interconnect. Of course, there is a countereffect when too many network users have to share limited resources

[13] Diewert (1986) defines infrastructure services to be public goods when all users have access to them at no charge.

[14] Robert Metcalfe founded 3Com Corporation and designed the Ethernet protocol for computer networks.