Payment System Efficiency and Pro-Competitive Regulation

Mats Bergman, Uppsala University[1]

mats.bergman@nek.uu.se

Preliminary work – for the Riksbank workshop, May 23-24 2003

May 5, 2003

1. Introduction

Competition is one mean for achieving efficiency within a payment system. However, competition can sometimes come into conflict with other possible means for achieving efficiency, such as large scale (scale economies), standardisation (network effects) and the firms' freedom to organise production. In addition, there is a tension between short-run competition and long-run competition. Maximum short-run competition sometimes comes at the expense of reduced incentives for long-run competition. A clear example is provided by patent rights and other intellectual property rights: if patent rights were not upheld, short-run competition would be intensified, while the incentives for long-run competition through innovation would be hampered.

Presently, there appears to be a trend towards a larger scale in payment systems, possibly resulting in many payment systems becoming national (or even regional) monopolies. Similarly, the European clearing and settlement industry is undergoing consolidation.[2] This suggests that an analysis of the possible benefits of pro-competitive regulation of payment systems is warranted. The general competition rules represent one type of such pro-competitive regulation. The infrastructure-access regulations that have been introduced in many deregulated industries represent another.

Historically, and as an alternative to pro-competitive regulation, traditional price regulation and government ownership have been widely used in industries where the owners of infrastructural monopolies have had the ability to exert market power. However, such regulator measures have decreased in popularity, as regulators and politicians have opted for deregulation. This does not preclude that in many (deregulated) industries, the competition rules are supplemented with pro-competitive sector-specific regulation. The clearest example is the access regulation applied to the telecom industry, but there exists other examples (e.g., within the electricity and airline industries).

[1] I am grateful to Gabriela Guibourg for valuable suggestions and comments.

[2] See the Giovannini Group's "Second Report on EU Clearing and Settlement Arrangements", April 2003.

The contrast between banking and telecom is interesting. In many countries, telecom started out as national monopolies.[3] Theoretical considerations, as well as the practical experience, suggested that relatively heavy regulatory measures were necessary, in order to "deregulate" the industry, i.e., in order to create a situation with effective competition. In contrast, the banking industry has always been competitive, although it was quite heavily regulated in the 1970s. During the 1970s and the 1980s, most of these regulations – rent regulations and credit regulations - were lifted in Sweden (and elsewhere) and around 1990, the conditions for obtaining bank charters (**concessions?**) were eased considerably. At the same time, banks and insurance companies were given the rights to enter into each other's markets. Hence, the banking deregulation was a proper deregulation, while telecom deregulation really meant that competition was introduced through a mixture of deregulation (easier entry) and the introduction of heavy-handed regulation (access regulation, i.e., price regulation of infrastructural services).

The contrast between banking and telecom becomes even more interesting, when one considers the trend towards competition in infrastructures in telecom and towards cooperation and concentration in infrastructure in banking. There is only one fixed telephony network in Sweden (as well as in most other countries), but there are currently three mobile telephony networks (GMS). After the introduction of UMTS, there will be one or two additional operators (3[4] and, possibly, Orange), two new national mobile networks (one owned by Telia and Tele2, one owned by Vodafone[5], 3 and, perhaps, Orange) and, in the largest cities, four new networks (owned by Vodafone, 3, Orange and Telia/Tele2). At the same time, the Bankomat and Minuten networks have become interconnected, and the compatability between Postgirot and Bankgirot has increased.

Currently, the European clearing and settlement system is consolidating, but international integration does not necessarily result in higher concentration at the national level and, therefore, does not give rise to the same immediate competition concerns. For example, the British competition authority, Office of Fair Trading, recently cleared the merger between the French-Belgian-Dutch and the British-Irish CSDs, CREST and Euroclear, respectively.[6] Still, it appears that now is the time to consider whether we want to introduce pro-competitive regulation. Experience, as well as theory, tells us that the regulations should be in place *before* private firms undertake significant investments in new infrastructure and *before* government-owned infrastructure is privatised.

Furthermore, there have already been serious controversies over access to existing payment infrastructures. In many cases, the concern has focused on the level of the interchange fee. This has been the case for card payments and for ATM transactions. At least at a superficial level, there is a clear parallel with the telecom industry, where many observers have pointed to high interconnection charges as responsible for the lack of genuine

[3] Actually, when the telecom networks were first built in the late 19th century and in the early 20th century, they were often privately owned and there were typically several telecom firms. R**eference!** Later on, they were nationalised in most European countries.
[4] Formerly Hi3G.
[5] The Swedish subsidiary of Vodafone previously operated under the name Europolitan.
[6] CSDs, or Central Securities Depositories, hold custody of securities and settle transactions. The Swedish VPC, Värdepapperscentralen, is a CSD. See OFT, 2002, "Proposed Acquisition by Euroclear plc of CRESTCo Limited", September.

competition. In addition, there have been controversies over the level of the fixed costs and the entry costs for small banks and new entrants in systems for card payment and ATM transactions, for national giro transactions and for clearing and settlement institutions.[7]

2. Scale economies v the benefit of competition

The above discussion appears to suggest that there is a trend towards fewer and larger payment systems. If this is correct, the efficient scale may be increasing, in a process that shifts the balance between the benefits from competition and the benefits from large scale.

The benefit of scale

A number of reasons suggest that, in general, there should be increasing returns to scale in production. Some fixed costs (e.g., management, R&D and computer systems) may not need to be duplicated, as the scale of production rises. Increased scale may allow a shift towards a more efficient technology (typically a more automated technology, with relatively higher fixed costs and lower variable costs).[8] A higher level of production will allow employees to become more specialised and will allow individuals and firms to move down the so-called learning curve. Finally, economies of massed reserves will allow firms to economise on production equipment, as random breakdowns or idiosyncratic demand and supply fluctuations will have less impact. However, these sources of scale economies will eventually peter out, and diseconomies will set in, such as increasing managerial costs due to the complexity of the operation, agency problems and, in many industries, transportation costs.[9]

On a general level, returns to scale in banking appear to be relatively modest. Wheelock and Wilson (2001), in a study on US banks, find significant returns to scale for banks with total assets below 300-500 million USD (approximately equivalent to the assets of some new entrant in the Swedish banking market, such as IKANO bank and ICA Banken, or a medium sized savings bank, such as Sparbanken Skaraborg). In addition, statistically insignificant point estimates suggest that there may be returns to scale up to perhaps 1 billion USD in assets.[10]

However, there exists empirical evidence that returns to scale are stronger in payment system, as reported by Bergendahl et al (2002); they find scale economies of 0.2 in a European cross-country study. This implies that costs increase with 2 % when volumes rise with 10 %. Similarly, Bauer reports average scale economies of 0.7-0.8 for check-processing offices, which, however, appear to be exhausted for the largest US offices.[11] Bauer and Ferrier (1996) report scale economies of approximately 0.5 for automated

[7] See, e.g., the EU Commissions press release IP/01/462, March 31, "Commission Raises Competition Concerns about Behaviour of Clearstream Banking AG". Clearstream is the German clearing and settlement institution. Preliminary, it has been found to having abused its dominant position by discriminating Euroclear.
[8] Cf. the so-called two-thirds rule, shown to apply for many chemical and metallurgical processes.
[9] See Tirole, 1988, section 1.2 and Scherer and Ross, 1990, chapter 4.
[10] See Wheelock and Wilson for further references.
[11] Quoted through Bergendahl et al, 2002.

clearing houses. In contrast, Felgran (1986) found that scale economies were insignificant for ATM networks with more than 1000 service points.

Network benefits

In addition to returns to scale from the supply (or cost) side, payment systems are characterised by returns to scale from the demand side. These are referred to as network effects. Network effects represent a special case of positive externalities. When an additional user connects to the network, this increases the utility for other connected users. The network effects can be direct, as in telephone networks, giro-payment systems and ACHs. They can also be indirect, as in payment-card systems and in the markets for computer software and hardware. If a larger number of consumers choose a particular payment card (e.g., Visa or a certain ATM card) or a particular type of office computer, there will be an incentive for more merchants to accept the payment card, more ATMs will be installed or there will be a larger supply of compatible software, respectively.[12]

In network markets, competition *between* networks must be distinguished from competition *within* systems. Examples of *inter*-network competition are PC computers v. Apple computers; American Express credit cards v. Visa v. Mastercard; and Swedish Postgiro v. Bankgiro. Examples of *intra*-network competition are competition between various PC producers; competition between acquirers within the Visa system; and competition between commercial banks that all provide giro solutions based on the Bankgiro. Note also that the distinction between intra-network competition and inter-network competition is not absolute, as there is often a degree of compatability even between supposedly non-compatible systems. The same EFTPOS terminals can be used for payment cards from several networks and customers can make giro payments from a Bankgiro account to a Postgiro account (and vice versa?).

Guibourg (2001) surveys the existing literature on network effects in ATM and ACH markets and provides an empirical study of network effects in the EFTPOS market. The general conclusion is that strong network effects and large returns to scale dominate over competition effects, in the sense that adoption rates are higher in markets with few competing networks – in fact often a single network - although there is normally *intra*-network competition. Guibourg reports that in 1999, the number of transaction per capita was ten times higher in countries with a single compatible EFTPOS network, than in countries with two or more incompatible networks. In addition, the growth rate in transactions per capita increased dramatically in countries where incompatible systems merged into one single compatible system. There appears to be less support for the existence of strong economies of scale in production, as the average number of proprietary (although often compatible) systems increases from three in 1988 to over seven in 1999. In addition, the three smallest countries in the sample are among the four countries with the

[12] With a higher level of demand, producers can benefit from returns to scale in production, offer more variety and offer higher service levels or a denser distribution network. A prerequisite for this to occur is that consumers choose compatible products. The four distinguishing features of network industries are 1) complementarity, compatability and standards, 2) switching costs and lock-in effects, 3) economies of scale in consumption or network externalities and 4) (significant) economies of scale in production. **Reference + more on external effect?**

highest number of transactions per capita, while the three largest countries are among the four countries with the lowest number of transactions per capita.

The benefit of competition

Just as fundamental as the economies of scale, are the benefits of competition. When competition is lacking, one or a few firms will possess market power, which in turn has four main adverse consequences. First, it will transfer welfare from consumers to producers.[13] Second, as the price rises above the competitive level, demand will fall below the optimal level – i.e., there will be allocative inefficiencies. Third, a low competitive pressure is generally believed to result in sub-optimal effort levels and X-inefficiencies. (I.e., weak cost control will result in too high costs.) Fourth, the existence of a monopoly profit may trigger socially costly lobbying for the favoured position, as well as other types of rent-seeking behaviour. Although regulation can ameliorate problems of the first and the second type, there is a substantial risk that it will not properly address problems of the third and fourth type. In addition, regulation brings new problems, such as regulatory risks (the risk that investment incentives *et cetera* will be reduced, because the regulator may be tempted to exploit the regulated firm after it has sunk the investment cost) and the direct costs of regulation.[14]

Firms have a strategic interest to overstate economies of scale and to downplay the benefits of competition. This is so, because a reduction in the number of competitors is typically beneficial for the industry and negative for consumers, while economies of scale will tend to benefit both categories. Hence, appealing to economies of scale provides a legitimate argument also in situations where the true rationale is a desire to reduce competition.

Some observers argue that the introduction of competition normally gives rise to cost savings and price reductions in the 25-75 % range (Winston, 1998, and a number of OECD studies, referred to in Gonenc and Nicoletti, 2000). However, based on an extensive review of the empirical literature on deregulation, Bergman (2002) arrives at the conclusion that a more realistic prospect is savings in the 5-10 % range. He draws a similar conclusion also in the section where studies of banking deregulation are surveyed.

3. The bottleneck problem – and possible remedies

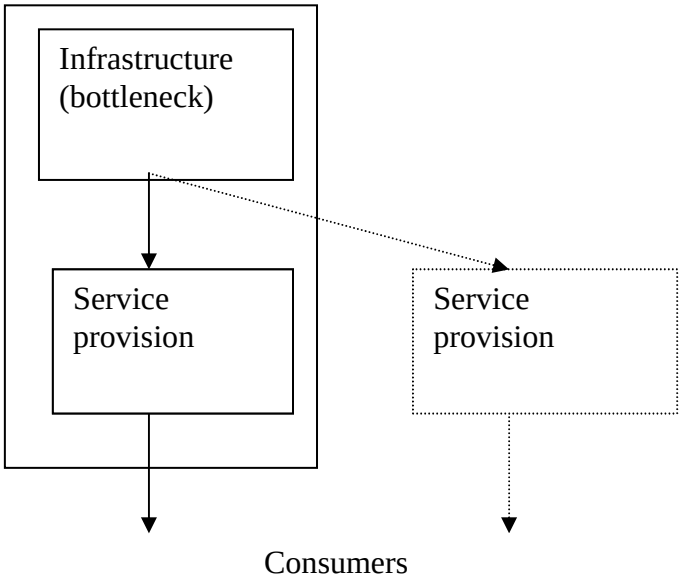
The fundamental bottleneck problem

Network industries typically have a vertical structure, with one (or a few) vertical stages where competition is non-viable or at least brittle, and one or more stages where

[13] Although, strictly speaking, this will only reduce welfare if we value consumer surplus higher than producer profit, welfare transfers from consumers to producers are normally considered to be negative.
[14] See Bergman, 2002.

competition is viable.[15] The situation is illustrated in Figure 1. By assumption, one firm only can be active in the upstream infrastructure market, the bottleneck, while several firms can be active in the downstream market – the market for service provision. For example, the upstream market can be establishing and maintaining a local telecom network or a payment system. The downstream market can then be the telecom services market or retail banking, respectively.

Figure 1. The bottleneck problem



By assumption, there will be market power in the bottleneck stage, which, in itself, gives rise to the negative consequences mentioned in the previous sub-section. However, control over the bottleneck can give rise to market power also in the potentially competitive downstream market.

The non-viability of competition in the bottleneck stage can be the result of economies of scale in production (which is typically the case for physical networks, like telecom, electricity and rail) or economies of scale in consumption – i.e., network effects. In the latter case, network effects may be so strong that "tipping" **(reference!)** will occur: as soon as one network gets an upper hand, all users will find it optimal to adapt the technology or products associated with that network. Examples of such tipping are the events that led the video system VHS to dominate over BetaMax; that allowed the US clearing and settlement institution DTCC to dominate completely the market for clearing by central

[15] An example of an industry with two distinct vertical stages where competition is non-viable is the electricity industry, where both (local) distribution and (long-distance) transmission of electric energy constitute "bottlenecks", while electricity generation and retail electricity sales are (potentially) competitive.

counterparties;[16] and that made Microsoft the almost-monopoly provider of operating systems for personal computers.

Although duplication of the bottleneck is not desirable, due to large scale economies or strong network effects, the creation of a monopoly will, if unchecked, normally result in welfare losses due to market power. The firm that controls the bottleneck is in a good position to extract the industry-monopoly profit. This can be achieved either by completely excluding competitors from access to the essential infrastructure, or by charging such a high price for infrastructural services that the monopoly profit for the industry as a whole accrues in the infrastructural stage alone.

In order to overcome these problems, various policies have been used in network industries. These policies can be divided into two broad categories – regulatory measures and structural measures – that can be employed in combination or separately.[17]

Regulatory measures

The traditional Swedish and European response of the bottleneck problem is government ownership. Postal services, telecom and rail services, for example, have been national monopolies in most EU countries. In banking, this method has been used less frequently, except for the central banks and some of their functions, such as settlements.

In the US, the preferred response to the bottleneck problem used to be regulation of consumer prices. In contrast to the situation in Europe, production was mainly in the hands of private firms. This method was used for some industries also in Europe – among these banking (rent regulation).

Instead of regulating consumer prices, it is possible to regulate the bottleneck price only (the price of the infrastructural service), i.e., access (price) regulation. If competition is viable in the non-bottleneck stages of production, access regulation should suffice to ensure effective competition.

Internationally and across many industries, there has been a move towards “deregulation”. Typically, market entry has been liberalised. In the US, firms have been given greater freedom over prices, as the consumer-price regulations have been lifted or eased. In Europe, however, price regulations have often been *introduced* during deregulation. This is so, since the government monopolies were often unregulated in a formal sense. In the early deregulations in Britain, privatisation was often accompanied by the introduction of consumer-price regulations. Later on and in other countries, there has been a tendency towards increased reliance on *access* regulation. This tendency is particularly pronounced in telecom. In banking, on the other hand, deregulations of entry and of consumer prices have rarely been followed by access regulation.

[16] See Oxera, 2003, Competing in Clearing and Settlement, Competing Ideas, April.
[17] Bergman, 2002, 2003.

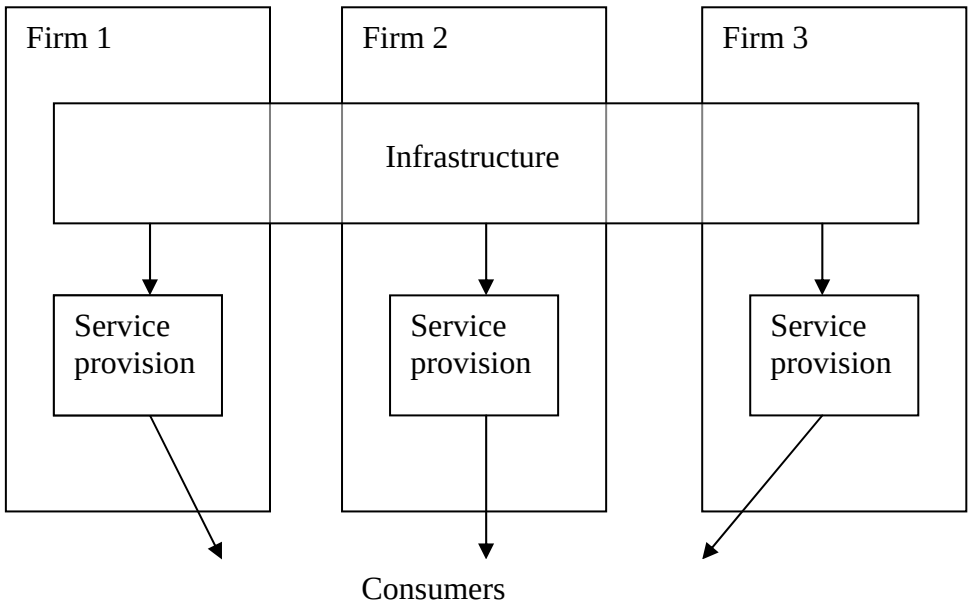
Structural measures

Vertical and horizontal separation have been proposed as structural measures for addressing the bottleneck problem. Vertical separation means that ownership of the bottleneck is separated from ownership of downstream (competitive) activities. The advantage is that the owner will not have incentives to favour one of the downstream operators. In contrast, a vertically integrated entity that faces competition downstream will often have an incentive to provide inferior infrastructural services, or price the services excessively high. The disadvantages with vertical separation are that vertical synergies cannot be fully exploited and that the problem of market power is not resolved. The latter implies that vertical separation must typically be combined with price regulation or government ownership.

Horizontal separation in itself will not resolve the bottleneck problem either. However, dividing one national bottleneck between several regional bottleneck operators can provide more information and may allow benchmarking between the operators. Hence, horizontal separation, as well as vertical separation, may facilitate regulation.

In banking, neither vertical, nor horizontal separation has been used very often (again, the national banks and their functions are the main exceptions). Instead, a third structural measure – infrastructural clubs – appears to be the preferred institution. An infrastructural club is an arrangement wherein firms that compete horizontally own the essential infrastructure jointly (see Fig. 2). Examples are the Visa and Mastercard systems, national giro systems like Bankgirot, and ATM networks. Examples from other industries are the airlines' computer reservation systems (CRSs), taxi switches and, sometimes, joint ownership of mobile-telephony infrastructure.

Figure 2. Infrastructural clubs



Infrastructural club provide some prospects for self-regulation. The owners/customers have a common interest of holding costs and prices low, while the firms may have strong incentives to compete for customers in the downstream market (e.g., retail-banking services). On the other hand, the common ownership of the infrastructure can conceivably be used to coordinate pricing in the downstream market, and there is a risk that large (incumbent) firms will not allow small (entrant) firms to join the clubs. In addition, conflicting views between the owners may increase transaction costs, which in turn may reduce efficiency and give rise to excessive inertia (lack of innovations).

The anti-competitive coordination risk can most easily be visualised using the analogy of patent pools.[18] Imagine that a number of firms compete horizontally in an industry. In order to be active in this industry, it is necessary to have access to certain technologies. These, in turn, are protected by patents. If each firm owns a patent that allows it to compete and if the number of firms is not too small, we will typically think that competition will be relatively fierce. However, if each firm sells its critical patent to a jointly owned patent pool, the situation will change drastically. The patent pool can now charge each firm with a high variable license fee. If this fee is set sufficiently high, the patent pool may earn the same profit as an industry-wide monopoly would, even if the firms compete fiercely for customers in the downstream market and earn no profit there. The profit that accrues in the patent pool can then be distributed between the firms in proportion to their respective owner shares of the pool. (If, on the other hand, profit is distributed in proportion to production, the firm will realise that the true marginal cost of using the infrastructure is not the nominal fee, but the nominal fee minus the average profit margin.)

4. Access regulation of payment systems

Patent-pool effects, discrimination and interchange fees

The analogy with patent pools suggests that infrastructural clubs may not be innocuous. Potentially, an infrastructural club could be used to achieve monopoly pricing just the same way as a patent pool can. This would be the case if the charges for infrastructural services were set at the monopoly level, and the accruing profit was then re-distributed in fixed proportions between the owners. As an example, the fees paid to Bankgirot could be set at such elevated levels that the banks would earn no profit in the downstream market. This would, potentially, give Bankgirot a handsome profit, which could be divided between the owners.

In practice, Bankgirot's fee structure follows the cost-plus principle and it is not likely that fees could be raised to the monopoly level, for a number of reasons. A drastic price increase, to the benefit of the owner banks, would probably be in conflict with competition law (see below). In addition, individual banks are not completely dependent on Bankgirot

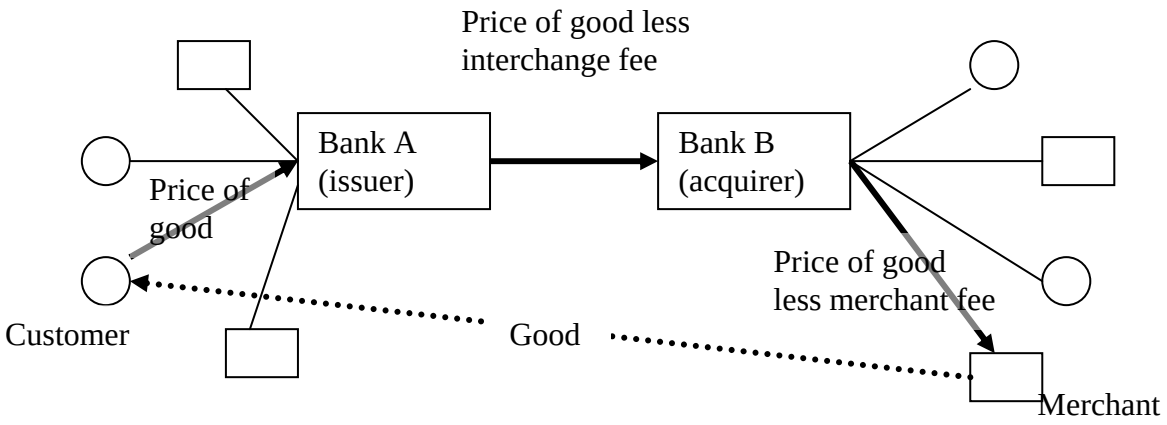
[18] See Priest, 1977.

for giro services. Postgirot (still) provides a certain competitive constraint for Bankgirot and there are other means of payment, e.g., checks, that could be used for distance payments. Hence, an individual bank would have incentives to buy too little services from Bankgirot, from the point of view of the collective of owner banks. In fact, there are no strong indications that huge profits are generated within the banking market's infrastructural clubs as such.

Even so, infrastructural clubs may tend to soften competition, since they transform fixed costs into variable costs. Typically, payment infrastructures are associated with large fixed costs, e.g., for computer hardware and software. If each bank acquires its own system, these costs will be seen as fixed and the banks' competitive actions will to a large extent be determined by the marginal costs, which are likely to be low. However, infrastructural clubs typically set fees close to average costs. From the perspective of an individual bank, these fees will be variable.[19]

The competition authorities' main concerns in regards to payment infrastructural clubs, however, have been focused on other issues. Internationally, the main concern has been that *horizontal* fees (interchange fees for banks or interconnection fees for mobile telephony operators) between the owners of the network components have been fixed at too high levels. (In contrast, note that in patent pools, the anti-competitive effect comes from *vertical* fees, paid to the pool as such.) The horizontal aspect of interchange fees is illustrated in Fig. 3.

Figure 3. Interchange fees



Often, banks are both issuers and acquirers. Depending on which customer buys from which merchant, either Bank A or Bank B is the acquirer. In the figure, a cardholder with a card from Bank A buys from a merchant whose transactions are acquired by Bank B, so for this transaction, Bank A is the issuer and Bank B is the acquirer. Following a purchase by

[19] Irrespective of how the cost is paid – as a fixed cost or as a variable cost – the consumer price must be set high enough for the firm to cover all costs. Hence, the question is whether a higher proportion of variable costs will result in higher or lower mark-ups above *average* costs. In a non-collusive duopoly, higher marginal costs will tend to increase prices. On the other hand, lower marginal costs may make it easier for the firms to collude, since the competitive alternative will be a more frightening alternative to collusion.

the cardholder from the merchant, the issuer deducts an amount corresponding to the purchase price from the cardholder's account (assuming a debit-card transaction). This amount, less the interchange fee, is transferred from the issuer to the acquirer. The acquirer, in turn, transfers the purchase price less the merchant fee to the merchant's account with the acquiring bank.

Within Sweden, the main concern has been that fees are set in such a way that new entrant and other small firms are discriminated against. This could, for example, be achieved by giving banks that use large quantities of the infrastructural services steep discounts, by having new banks pay high entry fees or by having all banks pay high annual fixed fees.

In summary, infrastructural clubs have given rise to the following competitive concerns:

- The owners can achieve coordination through a high vertical fee. Monopoly profits are generated in the bottleneck and then redistributed to the owners.
- The owners can achieve coordination through a high horizontal (multilateral) fee. This can give monopoly profits to acquirers/terminating operators.
- Joint ownership of infrastructure transforms fixed costs into variable costs, which softens competition.
- The infrastructural clubs are controlled by large firms, which have incentives to discrimination against small and joining members.
- The joint operation of assets allows coordination through other means, in particular through information sharing and through the creation of legal "links" between the firms that align their interests.

Joint determination of interchange fees

It may be thought that it is immaterial whether infrastructural charges are paid vertically, to an infrastructural-club entity, or horizontally, to competing firms that own infrastructure. For the individual bank, a vertical fee and a horizontal fee will be perceived alike – as a marginal cost of production. Hence, the initial intuition says that a multilateral interchange fee will have the same effect as a patent-pool license fee.

However, a horizontal fee will create incentives to compete for incoming fees, i.e., to compete for customers that in addition to paying themselves contribute to the firm's profit by generating incoming payments from other operators.[20]

In telecom, it has been claimed that the operators set too high interconnection fees. High interconnection fees raises the marginal cost of providing a telephone call, which in turn gives the operators incentives to increase the per-minute price of phone calls. On the other hand, as has been pointed out by Laffont, Rey and Tirole (1988a, b) (see also Laffont and Tirole, 2000), high interconnection fees raise the profit of the terminating operator. This gives the operators incentives to compete for customers who receive incoming calls

[20] This mechanism is not effective if the firms cannot compete for customers. An example may be agreements on interchange fees for international telephone calls and international mail delivery. For the latter, see the EU Commission's cases concerning the REIMS I and II agreements.

through other means than low calling prices, e.g., low monthly fees, handset subsidies or subsidies for receiving incoming calls.[21] Laffont, Ray and Tirole find that under certain conditions, a multilateral interconnection fee may still be welfare reducing and may still give rise to super-competitive profits. However, the effect is weaker and the conditions are stricter than the analogy with patent pools at first suggests.

Similarly, it has been claimed that banks set too high interchange fees. High interchange fees raise acquiring banks' marginal costs and, therefore, give incentives to raise the merchant fee.[22] On the other hand, issuing banks' profits increase, which gives banks stronger incentives to compete for cardholders. Annual fees and per-transaction fees charged to cardholders will fall.

A theoretical analysis of cooperation within payment-card associations

In an interesting recent paper, Rochet and Tirole (2002) analyse the welfare consequences of a multilateral interchange fee. Rochet and Tirole (RT) set up a two-sided market model of the card industry. Acquiring banks compete for merchants and issuing banks compete for cardholders. The two markets are interconnected, since cardholders use their cards with the merchants. In addition, acquiring banks pay a multilaterally determined interchange fee to the issuing banks. The two-sided nature of the market is the key difference between this market and a patent pool: if super-competitive prices are maintained on one market, the incentives to compete in the other market are strengthened. On the other hand, RT observe that earlier studies of the card industry has taken a naïve view on the banks' ability to exert market power over merchants. Previous studies had assumed that as soon as a merchant accepted cards, it had revealed itself to be better off with cards than it would have been without cards. In other words, merchants, as well as customers, adopted cards when the benefits of using cards exceeded the cost. However, as pointed out by RT, merchants could be compelled to accept the cards for *strategic* reasons. That is, the individual merchant will still accept cards only if this is to his benefit, but collectively, merchants may be worse off after cards are introduced.

The cause of the strategic effect is that a customer may decide *not* to purchase, upon finding that a merchant does not accept his preferred card. Hence, from the point of view of the individual merchant, there are two components that favour adopting the card. First, there are technological benefits (convenience, theft control et cetera), as recognised previously in the literature. Second, adopting the card increases demand, since some consumers will or cannot pay cash. Hence, from the point of view of an individual merchant, the product of his increased demand and his retail margin must be added to the technological benefits.

RT observe that at the social optimum, the sum of the merchant's and the cardholder's marginal benefits of paying with a card equals the sum of the issuer's and the acquirer's marginal costs. For example, this could be achieved if the merchant pays a fee equal to the

[21] If the calling pattern is "non-isotropic" **(?)** (i.e., if some subscribers receive more calls than they make and vice versa), there will in addition be incentives to compete for customers which receive many more calls than they make, such as costumer-support centres and internet-service providers **(?)**.
[22] The interchange fee constitutes approximately 80 % of the acquiring banks' costs. See below.

acquirer's marginal cost and the cardholder pays a fee equal to the issuer's marginal cost, but this is not the only possibility. Another possibility is that the merchant pays more than his marginal benefit and the cardholder pays less. RT assume that acquiring is competitive, while issuers have market power.[23] This suggests that in the absence of an interchange fee, cardholders would in market equilibrium have to pay a premium above the issuers' marginal cost, which would result in sub-optimal use of cards. Hence, it will be possible to increase welfare by introducing a positive interchange fee.

RT find that the interchange fee will be set at the highest level that is acceptable to the merchants. From a social point of view, this level may be too high. Then, there will be overprovision of cards. However, it may be that the socially optimal level of the interchange fee is higher than what is acceptable for the merchants. Then, the actual level of the fee will equal the constrained socially optimal fee.

The above results were derived under the assumption of a no-surcharge rule. If this rule is lifted, and under the assumption of no transactions costs, merchant resistance increases and there may instead be underprovision of cards. In conclusion, RT finds support for the idea that the introduction of a strictly positive MIF is likely to be welfare improving, but that the MIF may be set too high, if unrestricted. However, the outcome is sensitive to the assumptions made, e.g., that acquiring is competitive while issuing is not.

5. The role of competition law in payment-system markets

Competition law has three main pillars, or prohibitions. First, agreements between firms that reduce competition (cartels) are prohibited. Second, firms are not allowed to merge, if the merger will give the parties sufficient market power to significantly impede competition. Third, a "dominant" firm (a firm with significant market power) is not allowed to abuse its dominant position.[24]

The first two of these pillars can be seen as a system for weighing (and protecting) the benefits of competition against the benefits of large scale and of cooperation. Firms are often allowed to cooperate, but they are not allowed to enter into agreements that restrain competition unreasonably. In principle, the benefits of cooperation are weighed against cooperation's anti-competitive effects, although the weighing is not always done on a case-by-case basis.[25] Similarly, firms can merge unhindered by the law – unless the post-merger market share is too large – as low-market-share mergers are by default assumed to be beneficial, or at least harmless.

The third main pillar of competition law, the prohibition against abuse of dominance, prohibits some types of behaviour that has the effect or intention of throttling competition. For example, dominant firms are not allowed to discriminate other firms and there are

[23] If issuers were perfectly competitive, they would make no profit irrespective of the level of the interchange fee. Hence, the interchange fee would be indeterminate. See Schmalensee, 2002.
[24] The closest US correspondence is "monopolization".
[25] To simplify application of the law, there are "white lists", "black lists" and "group exemptions". In the US, black-listed behaviour is said to be prohibited "per se".

restrictions on their pricing and on their freedom to use rebate schemes. A special aspect of this prohibition is the "essential facilities doctrine". Under this doctrine, a dominant firm may sometimes be required to provide access to its infrastructure (or other "essential" facilities) at non-discriminatory prices.

Both EU's rules prohibiting anti-competitive agreements and the EU Merger Regulation are currently undergoing important changes. Under the proposed *modernisation* of the EU competition rules, firms will no longer be required to notify potentially anti-competitive agreements to the competition authorities. As a consequence, they will no longer be able to get legal exemptions from the prohibition.[26] The first aspect of the reform implies that the authorities will no longer automatically be informed of cooperative agreements concerning payment systems. The second aspect implies that any cooperative agreement will now have to be entered into at the parties' own risk. Although in theory the reform will not change the scope of the prohibition, it may have important practical consequences for the regulation of payment systems.

Under the proposed modification of the Merger regulation, less emphasis will be put on the post-merger market share and market-power level, while more emphasis will be put on the *change* in market power and the predicted consequences of that change. In addition, an "efficiency defence" will be introduced, suggesting that the merger's positive effects on efficiency shall be weighed against its negative effects.

In Sweden and elsewhere, competition law and its prohibition against anti-competitive agreements have been applied to payment systems. Within the financial industry, the merger regulation has most often been applied to bank mergers. However, two interesting merger cases which concerned payments systems are the Swedish case *Svenska Girot* and the merger between Euroclear and CREST, referred to above. In the following, I will focus on cases that have concerned the prohibition against anti-competitive agreements.

Swedish case law

In a sequence of cases, the Swedish Competition Authority evaluated the banks' cooperations concerning infrastructure for payment cards, ATMs and giro payments.[27] The Authority's analysis in these cases focused on the possible discrimination of small or entrant banks. There was a concern that the payment systems employed pricing schemes that were handicapping the smaller players in the market. For example, relatively high discounts were given to banks with a large number of transactions per year. In other instances, banks with few transactions had to pay surcharges; part of the payment systems' costs were covered through fixed annual fees, there were entry fees and in a series of cases concerning the jointly owned credit information agency *Upplysningscentralen*, the possibility that large profits might be accumulated and subsequently distributed to the

[26] An exemption is in this context an ex-ante decision ("förhandsbesked"), providing a legal guarantee that a certain agreement is allowed. Such ex-ante decisions have never been available in the US.
[27] See, e.g., case No. 1128/97, *FöreningsSparbanken AB (publ.) et al.*, (1999-10-29), concerning the Bankgiro and case No. 12/1999, *ABN Amro Bank N.V. et al.*, (1999-05-19), concerning "Dataclearingen", a jointly owned clearing institution.

owners of Upplysningscentralen were hinted at.[28] Effectively, this gave a higher price for non-owners.

One specific case dealt with the CEKAB, a jointly owned processor of electronic card-based payment transactions; ATM transactions as well as EFTPOS transactions.[29] CEKAB is 97 per cent owned by three of the four largest Swedish banks. The Competition Authority maintained that CEKAB's fee structure was discriminatory vis-à-vis smaller banks, which to a large extent were dependent on it. However, the Market Court found that on three grounds, the fee structure was acceptable.[30] First, the fees were cost based. Second, the fees did not have an appreciable effect. Third, commercially motivated discounting to a firm's largest customers may be acceptable, even in the absence of a cost justification.

After the Competition Authority lost the *Cekab* case in court, competition law has not been used very actively in order to facilitate small and entrant banks' access to payment systems in Sweden. However, it appears safe to conclude that in principle, competition law requires jointly owned infrastructural enterprises to grant small rivals access at non-discriminatory conditions.[31]

EU case law

Internationally, the prohibition against anti-competitive agreements has been used to challenge the rules of the international bank cooperatives Visa and Mastercard. In particular, the no-discrimination clause, the honour-all-card clause and the level of the interchange fee as such have been challenged.

The no-discrimination clause ensures that customers paying with cards are not surcharged, relative to customers paying cash. Visa International requires the individual banks to respect the no-discrimination clause in the banks' agreements with individual merchants. The European version of the clause prohibits any price differentials between customers paying cash and customers paying with cards. The US version of the clause only prohibits *surcharges* to card-paying customers; rebates to customers paying cash are accepted.[32] The no-discrimination clause was prohibited by the Swedish and the Dutch national competition authorities.[33] However, in a decision taken in 2001 by the EU Commission,[34] the no-discrimination clause was given negative clearance, as applied to *international* transactions. This is likely to have the implication that the clause is accepted throughout the Union, also for national transactions (also in Sweden and the Netherlands). The Commission's arguments for accepting the clause, as presented in the formal decision, were relatively poorly developed. The Commission primarily based its finding on the observation that relatively few merchants (5-10 %) used the option of surcharging in

[28] See cases No. 1124/93, 386/96, 861/97 and 851/2002, which granted UC individual exemptions.
[29] Swedish Competition Authority, case No. 605/1998.
[30] Case No. 1999:12, A 16/98, 1999-05-04.
[31] See Wetter at al., 2002, p. 226.
[32] Rochet and Tirole, 2002. Hence, the US rule is known as the "no-surcharge rule".
[33] Visa International only requires that the individual banks enforce the no-discrimination clause if the competent national authorities do not prohibit the application of the clause.
[34] Commission Decision of 9 August 2001, L 293/24, OJ 10.11.2001 (2001/782/EC).

Sweden and the Netherlands. An argument in favour of prohibiting the clause is that such a prohibition strengthens the bargaining power of the merchants vis-à-vis the banks. With the clause in force, merchants are in fact given a take-it-or-leave-it offer. Either they accept the card and do no surcharge, or they cannot accept the card at all. Without the clause, the merchant would have a third option: it could accept the card, but then surcharge customers that pay with cards. Presumably, this would give the merchants a better bargaining position. It may be that a strengthened bargaining position reduces the interchange fee even though, in equilibrium, relatively few merchants actually surcharge customers paying with cards.

The honour-all-cards rule requires a merchant that accepts, e.g., a Visa direct-debit card to also accept Visa deferred-debit cards and Visa credit cards. This is so, in spite of the fact that the merchant's fee can vary between cards of different types.[35] The Commission's argument for accepting this clause was that in its absence, customers would not be able to trust that a merchant accepted their card, even though the merchant purported to accept Visa cards. This, in turn, would endanger the universal acceptance of the system as a whole. In the US, merchants have instigated a class action that focus on the honour-all-cards rule.[36] Note also that the honour-all-cards issue is related to the no-discrimination issue: if merchants were happy with accepting direct-debit cards, but did not want to accept credit cards, they could surcharge customers using credit cards. Furthermore, it appears that much of the merchants' resistance, or all of it, would disappear if they could simply pass on the merchant fee to the final customers. They would have no reason to opt for prohibitively high surcharges.[37]

In a follow-up decision in 2002, the EU Commission gave a five-year individual exemption to the multilateral interchange fee (the MIF).[38] EuroCommerce, a retail, wholesale and international trade organisation, had complained that the MIF in fact amounted to horizontal price fixing, i.e., a cartel between the member banks. The fee was set by Visa EU Region and was applicable as the default interchange fee for cross-border transactions and, when no domestic default fee had been set, as the default interchange fee also for national transactions. Despite the MIF, every pair of banks was free to set another fee bilaterally. Since the introduction of the MIF in 1974, it had gradually increased.

The EU Commission found that the MIF restricted competition, but that it, on the other hand, contributed to the development of the Visa system and, therefore, potentially could be beneficial from a consumer-point-of-view. The Commission recognised the network aspects of the Visa system: cardholders benefit from a high number of merchants accepting

[35] The 2001 Decision, at 68.
[36] See below. Note that in the US, the merchants complain that the interchange fee is *the same* for off-line debit cards and for credit cards.
[37] In addition, the 2001 Decision dealt with rules that restricted cross-border issuing and acquiring, with territorial licensing and with the no-acquiring-without-issuing rule. The Commission imposed a partial liberalisation of cross-border activities, but it accepted the no-acquiring-without-issuing rule.
[38] Commission Decision of 24 July 2002, L 318/17, OJ 22.11.2002 (2002/914/EC). This means that the multilateral interchange fee was permitted for a period of five year. After that, the Visa would in theory have to re-apply for exemption. However, since the system of individual exemptions is being abolished, Visa will in practice not have to apply again. On the other hand, after the five-year-period, the Commission can itself initiate an investigation into the system and, at least in theory, prohibit the system or require modifications.

the card, while merchants benefit from a high number of cardholders. However, each type of consumer prefers that the other party bear the cost. I.e., cardholders prefer a high MIF, while merchants want a low MIF. The Commission identified maximum efficiency of the system with maximum size of the network, and argued that this would be achieved if each category of user paid a cost equal to the average marginal utility of that category.[39] However, as marginal utilities are difficult to measure, an "objective benchmark" (see below) for the system's cost would be an acceptable proxy for the marginal utility. Such a principle would ensure that each category got a "fair share" of the benefits provided by the system. From the Commission's analysis, it is clear that it perceived a risk that the MIF would be set too high. However, it recognised that it was necessary to establish a default MIF at *some* level, in order to reap the full benefits of the Visa system. In the absence of a MIF, there would be two possible outcomes. The first, and more likely, would be that the issuer had to charge the cardholder all its costs. This would imply a drastic rebalancing of the fee structure, since currently the MIF constitutes approximately 80 % of acquirers' costs. If the acquirers would not contribute to the issuers, cardholders' fees would increase significantly. This increase would result in too little usage of the system, since cardholder fees would be super-optimal and merchant fees would be sub-optimal. The second possible outcome would be bilaterally determined interchange fees. However, this would only be feasible in small national systems; Visa EU Region has 5000 members. Hence, the Commission found that a higher-than-zero MIF was conducive to maximum efficiency. Although the Commission did not try to pin down the exact optimal level, it argued that some objective criteria must be used to prevent the MIF from being set at a super-optimal level.

In order to get exemption from the competition rules, Visa agreed to lower the volume-weighted MIF for debit-card transactions by more than 50 %, to EUR 0,28, and to keep the MIF at that level for at least five years. In addition, Visa agreed to lower the average MIF applicable to credit and deferred-debit card transactions from approximately 0,85 % in 2002 to 0,7 % in 2007. Finally, Visa committed to set the MIF at a level that corresponded to the sum of three cost components, the "objective benchmark". The three components were the cost of processing transactions, the cost of the free funding period for cardholders and the cost of providing merchants with a "payment guarantee". In order to verify that the MIF reflected these costs, Visa committed to have accountancy firms make cost studies at regular intervals, and to present these to the Commission.

US case law

In the US, two high-profile on-going antitrust (competition law) cases have focused on the Visa and Mastercard systems.

The Department of Justice has challenged two features of the Visa and the Mastercard systems. First, that the banks in the board for each of the two systems participate in and have economic interests in the other system, the so-called duality. Second, that banks that

[39] As discussed above, this is a sufficient condition, but not a necessary condition.

issue Visa card or Mastercards are prohibited from issuing other general-purpose payment cards, such as American Express and Discovery Card.

On October 9, 2001, the U.S. District Court in the Southern District of New York upheld DOJ's charge on the second count, but not on the first. The decision has been appealed by Visa and Mastercard.[40]

A class action against Visa and Mastercard challenged the honour-all-cards clause. Roughly four million merchants sought $8.1 billion in damages. Under the US antitrust rules, the amount could potentially have been trebled to $24.3 billion.[41] However, a settlement was reached May 1 this year between a class of merchants and Visa and Mastercard. Visa and Mastercard agreed to pay $2 billion and $1 billion respectively. In addition, both agreed to significantly lower their fees to merchants. This is reportedly the largest settlement in antitrust history.

The key complaint was that the two card systems require that a merchant that accepts, e.g., a Visa credit card must also accept a Visa debit card. Hence, separate products are bundled together. In addition, the two systems have set the same interchange fee for off-line debit cards as for credit cards, despite large differences in costs between the two types of cards. According to the plaintiffs, this causes Visa's debit card interchange fee to be approximately 20 times higher than the interchange fee for a rival off-line debit card.[42] These two features, in combination, constituted the basis for the litigation.[43]

6. Discussion and conclusions

Clearly, forcing individual banks to set up proprietary non-compatible payment systems is not likely to be conducive to efficiency. However, if privately owned and operated payment systems are left unregulated, there are risks that anti-competitive effects will result in sub-optimal efficiency, as discussed above.

The general competition rules can to a certain extent prevent such anti-competitive effects, as indicated by the selective discussion of the competition case law of the payment-system industry. The strength of competition law is that, in principle, it can be applied to all types of anti-competitive behaviour. Hence, there is no need to envision in advance all possible means through which one or a few large firms can curb competition. Compared to sector-specific pro-competitive regulation it has, however, certain disadvantages. In general, it imposes less strict behavioural limitations than what can be achieved with sector-specific regulations. It will prevent dominant firms (or combinations of firms) from discriminating excessively against smaller rivals, but it may allow a certain degree of price differentiation.

[40] See http://www.usdoj.gov/atr/cases/f11700/11793.htm#COP
[41] The Wall Street Journal, April 2, 2003.
[42] See page 5 of the document http://www.inrevisacheck-mastermoneyantitrustlitigation.com/certification.pdf
[43] See Balto, 2000, for further details and an analysis.

It can impose access to natural-monopoly infrastructure ("essential facilities"), but the access price will typically be above average costs.[44]

For the above reason, competition law has in some industries been deemed to be insufficient for the purpose of protecting the interest of consumers. In such industries, access regulation complements the competition rules. The telecom industry is perhaps the industry that has been subject to the strictest set of regulations, while access regulation in the banking industry is relatively rare. A possible explanation for this divergence can be sought in the history of the two industries. The telecom industry has in most countries been a state-controlled monopoly, while banking has been a competitive or oligopolistic industry. While the telecom networks are typically owned by single firms, payment systems are typically infrastructural clubs. Mobile networks, however, have a similar ownership configuration as some card networks.

A possible regulatory strategy is to use competition law as a "first-line treatment". Sector-specific regulation will then only be considered if competition law cannot resolve the problem. Probably more incidentally than by calculation, this strategy has been used in some of the domestic Swedish deregulations.[45] However, at least when privatisation of a bottleneck is considered or when massive investments in new technologies are expected, it is likely to be less costly to implement a sector-specific regulation *before* the privatisation or the investments, respectively.

Theoretical analyses of the mobile telecom and the payment card industries suggests that even though the patent-pool effect does not translate fully into the multilateral interchange setting, there *is* a tendency that a multilateral interchange fee will be set too high. As concerns the telecom industry, the regulatory response has been to introduce sector-specific rules for the interchange fee (interconnection fee). In the payment-card industry, the regulatory initiatives have primarily been based on general competition rules.[46] The perceived success of telecom regulation suggests that there is a role for access regulation also in the payment-system industry.

However, it seems appropriate to re-iterate Laffont and Tirole's (2000) warning against drawing analogies between different network markets:

---

[44] As an illustrative example, the Swedish Competition Authority was able to reduce the incumbent telecom operator's (Telia's) interconnection fee from 0.35 SEK to 0.215 SEK for single-segment (local) access. Later on, regulatory decisions based on the telecom legislation reduced the fee to 0.069 SEK.

[45] Bergman, 2002.

[46] An interesting exception is the EU Commission's newly-introduced regulation of the *consumer* price of international payments. According to Regulation 2560/2001, the cost of international Euro-denominated payment transactions within the Union should be the same as that of intranational payments.
Perhaps even more interesting is the Australian legislation concerning payment systems. On its homepage, the Australian central bank writes that "[t]he Payments System Board (PSB) of the Reserve Bank oversees the payments system in Australia. The PSB is responsible for promoting the safety and efficiency of the payments system in Australia. Through the Payment Systems (Regulation) Act 1998 (PSRA), and the Payment Systems and Netting Act 1998 the Reserve Bank has one of the clearest and strongest mandates in the world to oversee the operation of the payments system."

One should be careful, though, before importing lessons drawn from one industry into another, since networks can differ substantially.[47]

Before regulations are introduced, however, it is also worth recollecting the trade-off between, on the one hand, short-run competition and, on the other hand, the incentives for investments and long-run competition. Too strict regulations will clearly restrict the incentives to invest and to engage facilities-based competition.

References

Balto, David A., 2000, Creating a Payment System Network: The Tie that Binds or an Honorable Peace?, *The Business Lawyer*, 55, 1391-1408.

Bauer, Paul and Gary Ferrier, 1996, Scale Economies, Cost Efficiency, and Technological Change in Federal Reserve Payment Processing, *Journal of Money, Credit and Banking*, 28, 1004-39.

Bergendahl, Göran, David Humphrey, Ted Lindblom and Magnus Willesson, 2002, *What Does it Cost to Make a Payment?* Mimeo.

Bergman, Mats, 2002, *Lärobok för regelnissar – en ESO-rapport om regelhantering vid avreglering*, Ds 2002:21, Stockholm. (A Textbook on Deregulation. Experiences from six Network Industries.)

Bergman, Mats, 2003, Competition Law, Competition Policy, and Deregulation forthcoming in *Swedish Economic Policy Review*.

Felgran, Steven D., 1985, From ATM to POS Networks: Branching, Access, and Pricing, *New England Economic Review*, 44-61.

Gonec, Rauf and Giuseppe Nicoletti, 2000, *Regulation, Market Structure and Performance in Air Passenger Transportation*, OECD, Economics Department Working Paper No. 254.

Guibourg, Gabriela, 2001, *Interoperability and Network Externalities in Electronic Payments*, Sveriges Riksbank Working Paper Series, No. 126, Stockholm.

Laffont, Jean, Patrick Rey and Jean-Jacques Laffont, 1998a, Network Competition: I. Overview and Nondiscriminatory Pricing, *Rand Journal of Economics*, 29, 1-37.

Laffont, Jean, Patrick Rey and Jean-Jacques Laffont, 1998b, Network Competition: II. Price Discrimination, *Rand Journal of Economics*, 29, 38-56.

[47] Page 181. It appears that the same warning can be applied to a specific theoretical model of *one* network industry.

Laffont, Jean and Jean-Jacques Laffont, 2000, *Competition in Telecommunications*, MIT Press, Cambridge, MA.

Priest, G.L., 1977, Cartels and Patent License Agreements, *Journal of Law and Economics*, 20, 309-377.

Rochet, Jean-Charles and Jean Tirole, 2002, Cooperation Among Competitors: Some Economics of Payment Card Associations, *Rand Journal of Economics*, 33, 549-570.

Scherer, Fredrick M. and David Ross, 1990, *Industrial Market Structure and Economic Performance*, Houghton Mifflin, Boston, MA.

Schmalensee, R, 2002, Payment Systems and Interchange Fees, *Journal of Industrial Economics*, 50, 103-122.

Tirole, Jean, 1988, *The Theory of Industrial Organization*, MIT Press, Cambridge, MA.

Wetter, Carl, Johan Karlsson, Olle Rislund and Marie Östman, 2002, *Konkurrenslagen – en handbok*, Thomson Fakta, Stockholm.

Wheelock, David C. and Paul W. Wilson, 2001, New Evidence on Returns to Scale and Product Mix among U.S. Commercial Banks, *Journal of Monetary Economics*, 47, 653-674.

Winston, Clifford, 1998, U.S. Industry Adjustment to Economic Deregulation, *Journal of Economic Perspectives*, 12, 89-110.