The Theory of Interchange Fees: A Synthesis of Recent Contributions

JEAN-CHARLES ROCHET*

Toulouse University, IDEI and GREMAQ

Abstract

We synthesize the results of the recent theoretical literature on the determination of interchange fees by payment card associations. We analyze in particular the conditions under which these interchange fees are excessively high, as compared with social optimum. These conditions involve several parameters: the intensity of competition between banks and between merchants; the elasticities of demand on both sides of the market, as well as the degrees of heterogeneity among merchants and among cardholders. A crucial element is the competitive edge that merchants can gain by accepting cards.

1 Introduction

Interchange fees are a crucial determinant of price structure and transaction volumes in open payment card networks. Following Baxter (1983), an important antitrust literature[1] has discussed the potential anticompetitive effects of the collective determination of interchange fees within payment card associations. However, no systematic theoretical analysis of price determination in payment card networks was available. The situation has changed in recent years: several formal models of the payment card industry have been developed, allowing a more rigorous analysis of the impact of interchange fees on prices and volumes of activity in payment card networks. These models have also highlighted the existence of common patterns between this industry and other network industries like Internet, media, video games and software, which have been termed “two-sided” markets.[2] The objective of this paper is to synthesize the findings of this recent theoretical literature[3] on interchange fees and suggest further directions of research.

* Mailing address: Pl. Anatole France - 31042 Toulouse Cedex, France. E-mail: rochet@cict.fr This article is an outcome of a long term research program with Jean Tirole, who played a crucial part in the development of the ideas presented here. I also thank Yossi Spiegel, Julian Wright, and an anonymous referee for their detailed comments. All errors are mine.

[1] See in particular Carlton and Frankel (1995), Evans and Schmalensee (1995), Frankel (1998), Chang and Evans (2000), and Balto (2000).

[2] See Rochet and Tirole (2003), Armstrong (2002) and Evans (2002).

[3] Schmalensee (2000b) provides a non technical survey of the anti-trust literature on interchange fees. By contrast, the present paper focuses on economic modeling.

2 A fundamental externality

The choice of a payment instrument in any transaction involves a fundamental externality, since it affects the costs and benefits of both parties to the transaction. For example, if the buyer insists on paying by cash (which is legal tender), the seller will incur the cost of handling and holding the cash until it can be deposited in a bank.[4] On the other hand, a cash payment allows the seller to save the fees charged by his bank for managing card payments. Symmetrically, if a seller refuses a payment by card or check, he typically forces the buyer to incur the cost of finding an ATM and withdrawing cash. It may also prevent the buyer from receiving the benefits that are often associated with a card payment (such as deferred debit or free interest period, frequent flyers miles or cash back bonuses).

Payment card networks are also characterized by a more classical network externality. Indeed when a seller decides to join a payment card network (and thus implicitly commits to accepting the cards issued by the members of this network), this increases the potential utility of buyers who hold such cards, by offering them a new opportunity for using their cards. This is similar to the positive externality generated when a new user joins a telecom network. This network externality becomes less and less important as the network matures, when virtually all potential users have joined.[5] By contrast, even in a mature network (where virtually all buyers hold cards and sellers accept them), the fundamental externality identified above remains important: the choice of the payment instrument is ultimately a decision of the buyer, that impacts the net costs of the seller. Our objective in this section is to study the consequences of this usage externality.

To simplify the analysis, we start with the case where the number of transactions is fixed and where transactions can only be paid for either by using a card or cash. Thus, the efficiency of card usage is determined by answering a unique question: which transactions are settled using a card rather than cash?

The answer to this question depends on the difference in the net utilities accruing to buyers and sellers for a card payment and for a cash payment. Let us denote by b^B this difference in utility for the buyer and b^S for the seller, for a typical transaction.[6] Similarly, c denotes the total cost of a card payment for the two banks who provide the payment service:[7] the bank of the buyer, called the issuer, and the bank of the seller, called the acquirer. Since a cash payment is costless for the banks,[8] c can also be interpreted as the incremental cost of card versus cash.[9] Thus social welfare is maximized (i.e. the use of cards is socially efficient) whenever card payments occur if and only if:

(1) $b^B + b^S \geq c.$

[4] For simplicity, we will call these parties the buyer and the seller, instead of using the technical terms: payor and payee.

[5] For a dynamic analysis of network externalities see Katz and Shapiro (1985, 1986).

[6] A crucial element of our analysis will be the heterogeneity of b^B (across buyers) and b^S (across sellers). In our first model (Section 3) only b^B is heterogenous. In the second model (Section 5) both buyers and sellers are heterogenous. It is also important to distinguish ex-ante heterogeneity (across types of buyers and sellers) and ex-post heterogeneity (across types of transaction).

[7] c may also include the marginal cost of the card network, but we neglect it in this analysis.

[8] We are referring here to direct costs only. Cash payments are indirectly costly to banks, who have to serve ATMs with cash. However they also receive fees for cash withdrawals.

[9] This would not be true if the alternative payment instrument was a check instead of cash (see footnote 10).

Baxter (1983) was the first to emphasize that perfect competition between banks does not on its own lead to this condition (i.e. to an efficient choice of payment instruments). Indeed, perfect competition without transfers implies that p^B (the unit price of card services for buyers) equals c^B (the marginal cost of issuers) and similarly that p^S (the unit price of card services for sellers) equals c^S (the marginal cost of acquirers). Notice that $c^S + c^B = c$. In the absence of side-payments and strategic considerations by sellers (both aspects are studied later), a card payment will take place if and only if both parties agree; that is:

(2) $b^B \geq p^B = c^B$ and $b^S \geq p^S = c^S$.

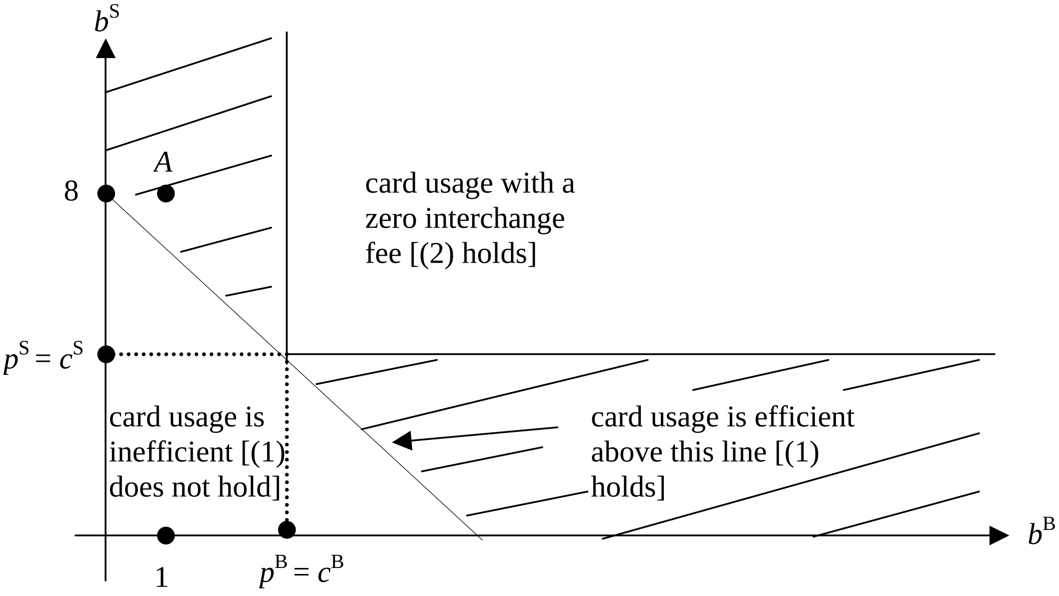
b^S
A
8
card usage with a zero interchange fee [(2) holds]
$p^S = c^S$
card usage is inefficient [(1) does not hold]
card usage is efficient above this line [(1) holds]
b^B
1
$p^B = c^B$

Figure 1: Under-efficient usage of card services when the interchange fee is zero. The shaded area represents foregone card transactions that would have been efficient

As illustrated by Figure 1, condition (2) is more restrictive than condition (1), which means that, in the absence of side payments, there will be under efficient usage of card services, even if banks are perfectly competitive.[10] This is due to the externality described above: consider for example the case where $b^B = 1$, $b^S = 8$, $c^B = 3$, $c^S = 2$ (this corresponds to point A in Figure 1). By not using the card (since $b^B < p^B = c^B$) the buyer inflicts a

[10] This is even more true if the alternative payment instrument is a check, of total cost $c_0 > c$. The incremental cost of cards becomes negative, whereas checks are often free of charge for users. In this case the condition for efficient usage of cards becomes $b^B + b^S \geq c - c_0$, whereas in the competitive equilibrium without interchange fees, the condition for cards usage remains $\{b^B \geq c^B$ and $b^S \geq c^S\}$(this is because check payments are not charged by banks). If checks were charged at their unit costs, say c_0^B and c_0^S , the reasoning of the text would apply, provided that the costs of card payments are replaced by the incremental costs of card versus checks. Notice that in this case, an interchange fee on checks could be enough to restore efficiency. However, in practice more than two payment instruments are available, which complicates the analysis. In particular, an efficient use of payment modes requires that interchange fees be set appropriately for all non-cash instruments.

negative externality ($c^S - b^S = -6$) on the seller, and prevents a socially efficient card payment (since $b^B + b^S = 9 > c^B + c^S = 5$).

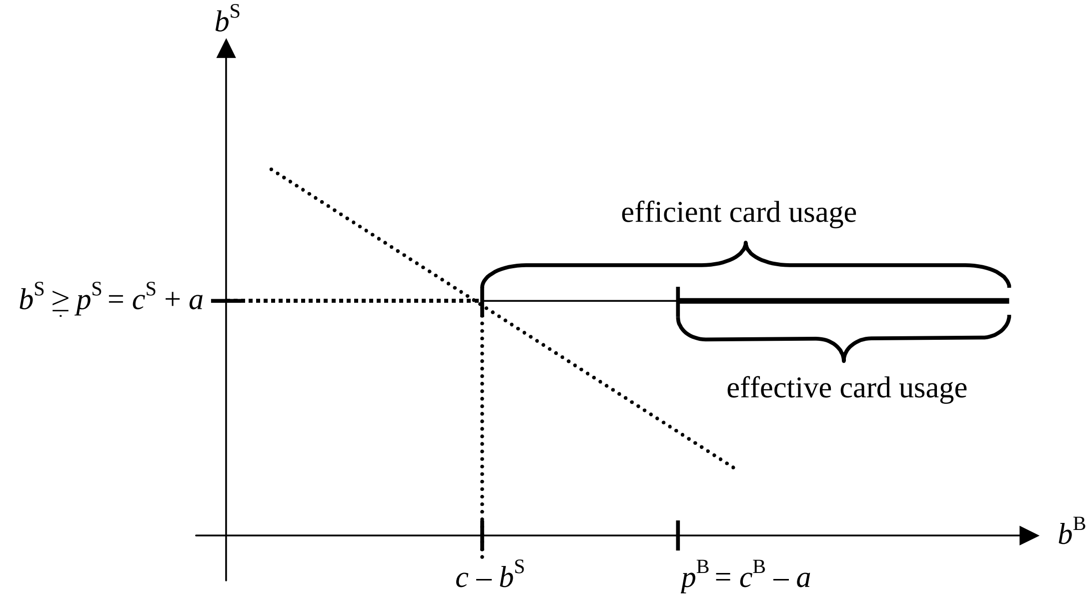

Figure 2: Card usage with homogenous sellers and a too low interchange fee $a \leq a_0 = b^S - c^S$

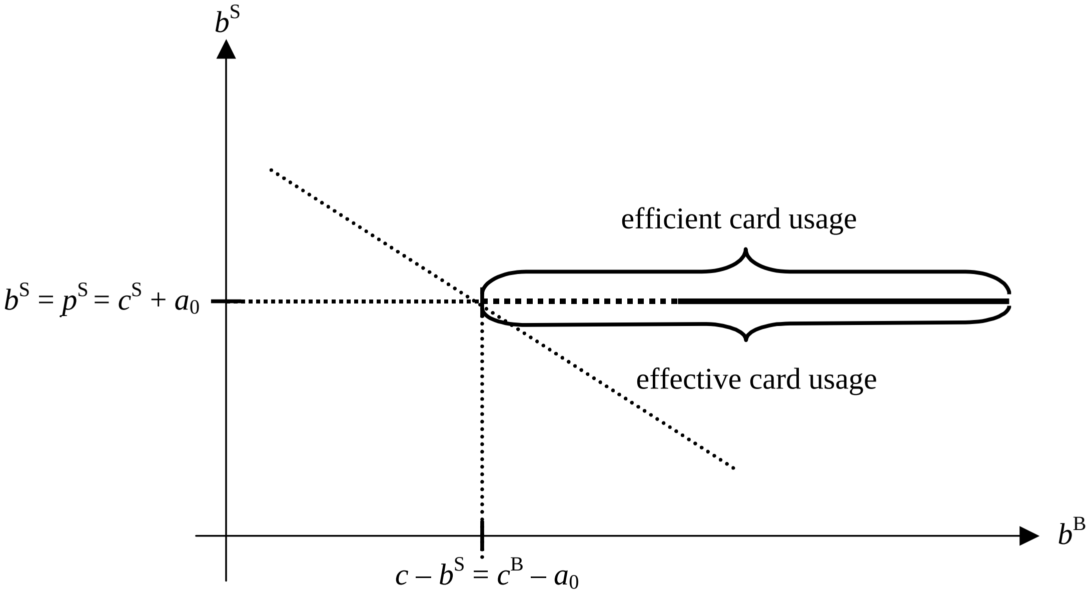

Figure 3: Card usage is efficient when $a = a_0$ (Baxter's interchange fee)

In the case where sellers are homogenous (i.e. b^S is the same for all sellers) and banks are perfectly competitive, Baxter (1983) shows that this inefficiency can be corrected if the acquirer pays an interchange fee of $a_0 = b^S - c^S$ to the issuer. The net cost of the issuer becomes $c^B + c^S - b^S = c - b^S$. Since there is perfect competition between banks, this change in the issuer's cost is fully passed through to buyers (p^B becomes $c - b^S$) and the card transaction takes place if and only if:

(3) $b^B \geq c - b^S$,

which is equivalent to (1), the social efficiency condition. In a perfectly competitive context, an interchange fee set at the optimal level a_0 allows the internalization of the fundamental externality described above and restores efficiency of card usage.[11,12] This is illustrated by Figures 2 and 3 above.

In their criticism of Baxter's analysis, Carlton and Frankel (1995) argue that this externality can also be internalized by a different method, namely allowing sellers to charge differentiated prices for cash and card payments. Indeed, if sellers themselves are perfectly competitive, they fully pass through their net cost $c^S - b^S$ of card payments to buyers, who then use their card if and only if

(4) $b^B - p^B \geq c^S - b^S$,

In the case of competitive issuers ($p^B = c^B$) this also leads to the efficiency condition:

$b^B + b^S \geq c^B + c^S = c.$

In practice however,[13] differentiated pricing for cash and card payments is seldom used by sellers (this phenomenon is called "price coherence" by Frankel, 1998), probably because of transaction costs.[14] But if the majority of sellers do not surcharge buyers, the volume of payments (which depends both on the proportion of sellers who accept cards and on the proportion of buyers who are willing to pay by cards) is a function of both prices p^B and p^S and not only on their sum. In this case, the level of the interchange fee matters.

Then a natural question arises: do credit card networks choose an appropriate level of the interchange fee or is there a systematic bias toward under- or over-provision of card services? Unfortunately, Baxter's framework cannot provide an answer to this question. This is because in a perfectly competitive world, banks make no profit regardless of the level of the interchange fee. It is therefore impossible to predict how the payment card

[11] If b^S varies across sellers but is observable, Baxter's result remains valid if the network sets differentiated interchange fees.

[12] Notice that no network effect is involved in this example. Network effects only become important when fixed fees and fixed costs are introduced, or when one of the parties (typically the seller) has to commit ex-ante to accept cards. In this case, his acceptance decision is influenced by the number of cardholders.

[13] This is acknowledged by Carlton and Frankel.

[14] It is true that payment card networks try to discourage differential pricing. In some countries, they have been able to impose uniform pricing (this is called the No Discrimination Rule) on merchants. In the U.S., surcharges for card payments are forbidden (at least in some states: see Chakravorti and Shah, 2001 for a detailed discussion of surcharges and discounts in the U.S.) but cash discounts are allowed. In the U.K. any form of restriction is illegal, yet differentiated pricing is not frequently observed.

network will select its interchange fees. A second potential criticism of Baxter's analysis is that sellers' private and common interests may diverge: sellers who accept cards are likely to attract more customers, which affects their acceptance decision, a consideration that is absent in Baxter's analysis. Finally, banking industries are far from being perfectly competitive in many countries. Carlton and Frankel (1995) have expressed the concern that interchange fees could be used as collusive devices when banks have market power. To study these questions, it is therefore important to extend Baxter's analysis to imperfectly competitive banking industries. This is the topic of the next section.

3 A first model of the payment card industry

Rochet and Tirole (2002) provided the first fully-fledged model of an imperfectly competitive payment card industry, allowing a comparison between privately optimal and socially optimal interchange fees. With respect to Baxter's analysis, two important features of the payment card industry are added: imperfect competition between issuers and strategic behavior of sellers. For tractability, Rochet and Tirole make some simplifying assumptions: there is perfect competition among acquirers, and the total number of transactions (cash plus card) is fixed, and normalized to one. The first assumption seems to be a good approximation of the U.S. situation but may be less acceptable for other countries (it is relaxed in Section 5).[15] The second assumption amounts to focusing on the choice of payment instruments for a given volume of transactions and therefore neglecting the impact of price variations in payment services on the demand for final goods and services: it also seems reasonable in a first step.[16] Finally, Rochet and Tirole assume for simplicity that sellers are homogenous. This assumption is relaxed later.

A crucial element in the analysis of Rochet and Tirole (2002) is capturing the impact of imperfect competition on issuers' pricing policy. Rochet and Tirole do not rely on a specific model of imperfect competition but they just assume that the equilibrium price in the issuers' market equals some function $p^B = f(\gamma^B)$ of the net cost $\gamma^B = c^B - a$ of issuers. Due to imperfect competition, $f(\gamma) > \gamma$, so that issuers make a positive profit.[17] Rochet and Tirole make the plausible assumption that $0 \leq \dot{f} \leq 1$ (throughout the remainder of the paper, dots denote derivatives). This assumption means that when the cost γ^B of issuers increases, the price $p^B = f(\gamma^B)$ of card payments (for buyers) also increases (this is because $\dot{f} \geq 0$) but issuers margins $p^B - \gamma^B$ do not (this is because $\dot{f} \leq 1$). In other words, issuers do not increase prices more than their costs increase. This assumption[18] implies that an increase in the interchange fee a (which decreases issuers' costs) has a positive impact on buyers' demand for card payments (since p^B decreases) and also on issuers' margins. Therefore issuers' profit increases with a.

The second key element of Rochet and Tirole's model is the modeling of sellers' strategic behavior in their decisions to accepting or refusing a card. Instead of considering,

[15] Evans and Schmalensee (1999) argue that issuing is typically less competitive than acquiring since brands and consumer loyalty are important for issuing but not for acquiring. On the other hand, acquiring is very competitive in some countries (e.g. the U.S.) but very concentrated in other countries (e.g. Israel).

[16] Two papers by Schwartz and Vincent (2000, 2001) relax this assumption in a particular context.

[17] The limit case $f(\gamma) = \gamma$ (perfect competition) is also considered by Rochet and Tirole (2002).

[18] This assumption is very natural: it is satisfied in virtually all models of imperfect competition that are used in the Industrial Organization literature. It is also empirically plausible.

as Baxter did, that sellers only look at their costs and neglect the competitive edge that they can obtain by accepting a card, Rochet and Tirole capture this effect by using a Hotelling type of model where sellers compete not only in prices but also in the quality of their services: card acceptance increases the quality of service (to cardholders) for a given price. Card acceptance is then motivated not only by the net savings in payment costs, but also by a search for a better competitive position.

More specifically, the timing of strategic decisions in Rochet and Tirole (2002) is the following:

- The level a of the interchange fee is set (either by the payment card network or by a regulator).

- The members of the network compete in prices (imperfectly on the issuers' side, perfectly on the acquirers' side), which results in equilibrium prices for card payments: $p^B = f(c^B - a)$ for buyers, $p^S = c^S + a$ for sellers.

- Sellers compete in prices and service qualities, which results in equilibrium prices for final goods or services and decisions to accept or refuse card payments. Differentiated pricing for card and cash payments is ruled out in this section.

- Buyers observe sellers' prices and, for a fraction α of them, also observe which sellers accept cards. Then buyers select which store they patronize and whether or not they pay by card.[19]

Rochet and Tirole (2002) show that the maximum price that sellers are ready to pay for card services is higher than in Baxter (1983). It becomes $b^S + \alpha v^B(p^B)$,[20] where $\alpha \in [0,1]$ denotes the proportion of buyers who are informed (about which sellers accept the card) before they select a store, and $v^B(p^B)$ is the average net surplus generated by card transactions for the buyers who pay by card.[21] The reason why sellers' willingness to pay for card services is higher than in Baxter's case (where the maximum acceptable price by sellers was b^S, which corresponds to $\alpha = 0$) is that sellers are afraid to lose customers if they refuse the card.

Since acquirers are supposed to be perfectly competitive, the price of payment services for sellers equals acquirers' net marginal cost:

[19] In Rochet and Tirole (2002), p^B is supposed to be a fixed fee, paid ex-ante by cardholders. This implies that some buyers do not hold cards. In conformity with the subsequent literature (e.g. Wright, 2001, and Guthrie and Wright, 2003), we assume instead that there is no fixed fee and that p^B is only paid when a card transaction occurs. This simplifies the analysis: all buyers hold cards but use them only when $b^B \geq p^B$. Notice however an asymmetry between buyers and sellers: when sellers accept cards, the payment mode is chosen by buyers. Thus sellers have to be given incentives to accept cards (network externality) whereas buyers have to be given incentives to use cards (usage externality).

[20] Wright (2003) obtains the same upper bound in a model where sellers compete *à la* Cournot. He assumes that b^B is only observed by buyers after entering the store (which means that unobservable heterogeneity of b^B is transaction specific and not buyer specific). Under this assumption, the upper bound found by Rochet and Tirole (2002) holds for classical models of competition between sellers (Hotelling, Cournot, differentiated Bertrand etc).

[21] Specifically, $v^B(p^B) = E[(b^B - p^B)|\, b^B \geq p^B]$. In the original formulation of Rochet and Tirole (2002), v^B is replaced by $v^B + p^B$ in (5), because p^B is supposed to be a fixed fee, already incurred by buyers when they come to the shop.

$p^S = c^S + a$.

Thus, cards will be accepted by merchants whenever:

$c^S + a \leq b^S + \alpha v^B[f(c^B - a)]$,

which is equivalent to:

(5) $a - \alpha v^B[f(c^B - a)] \leq a_0 = b^S - c^S$.

Lemma 1[22] *Cards are accepted by merchants if and only if the interchange fee is less than some threshold $\bar{a}$. This threshold $\bar{a}$ is defined implicitly by the equality in (5).*

$\bar{a}$ is the maximum value of the interchange fee (henceforth IF) that is compatible with sellers' acceptance. Given that issuers' profit increase in a, $\bar{a}$ is the privately optimal level of the IF (i.e. the one set by the network).[23] This is represented in Figure 4.

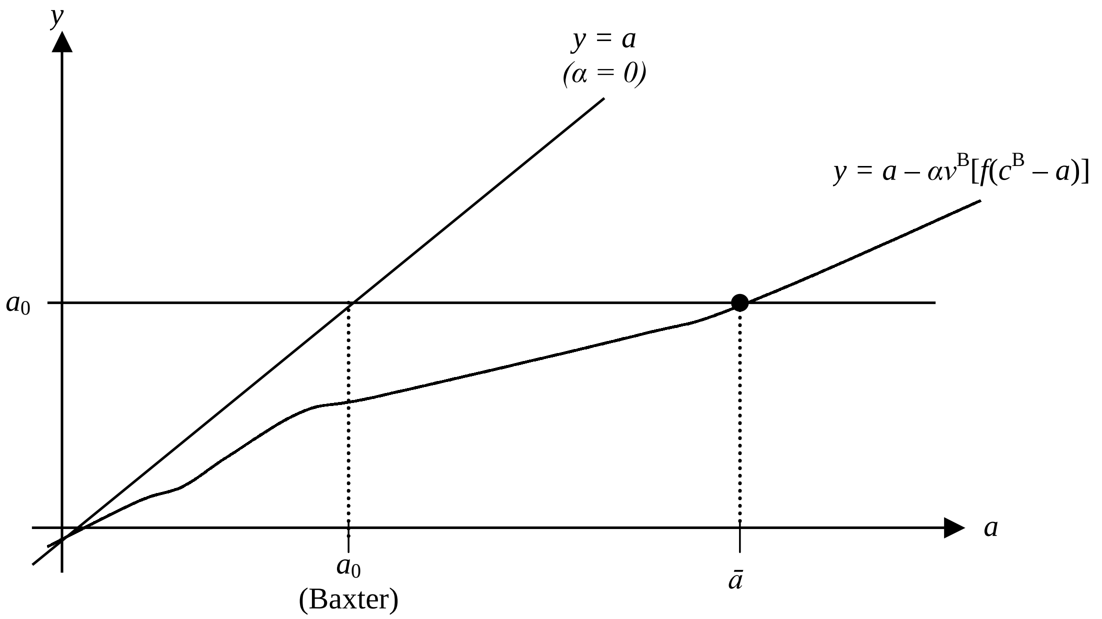

Figure 4: Determination of $\bar{a}$, the privately optimal interchange fee in Rochet and Tirole (2002) (maximum level compatible with sellers' acceptance)

Proposition 1 *The privately optimal IF equals the maximum value of the interchange fee $\bar{a}$ that is compatible with sellers' accepting cards. It is defined implicitly by:*

(6) $\bar{a} - \alpha v^B[f(c^B - \bar{a})] = a_0 = b^S - c^S$.

[22] Proofs of all lemmas and propositions are in the Appendix.
[23] Since acquirers are perfectly competitive, the total profit of the network's members is equal to issuers' profit.

Notice that $\bar{a}$ increases with α, and is equal to Baxter's interchange fee $a_0 = b^S - c^S$ when $\alpha = 0$. Thus as soon as $\alpha > 0$, the IF chosen by the association is higher than the one predicted by Baxter (i.e., $\bar{a} > a_0$), which can be explained by the fact that Baxter's analysis underestimates the sellers' willingness to pay for cards.

By contrast, the socially optimal level a^* of the interchange fee is obtained by maximizing social welfare under the constraint that merchants accept the card. Let us denote by $h(b^B)$ the density of the statistical distribution of buyers' incremental utilities. As a function of $p^B = f(c^B - a)$, social welfare equals:

(7) $W(p^B) = \int_{p^B}^{+\infty} (b^B + b^S - c)h(b^B)db^B$.

The constraint that sellers accept the card ($a \leq \bar{a}$) amounts to saying that the buyers' price for card services ($p^B = f(c^B - a)$) cannot be smaller than $\bar{p}^B = f(c^B - a)$ (remember that $f(\cdot)$ is increasing). Thus, the welfare function $W(p^B)$ has to be maximized under the constraint $p^B \geq \bar{p}^B = f(c^B - \bar{a})$. By differentiating (7), we obtain:

$$\dot{W}(p^B) = -(p^B + b^S - c)h(p^B).$$

Thus, if the constraint is not binding, the maximum of W is obtained for $p^B = c - b^S$, which corresponds to the social value of the externality imposed by the buyer (to the rest of the economy, i.e. the issuer, the acquirer and the seller) when he pays by card. The associated interchange fee a^{**} is given by

(8) $c - b^S = f(c^B - a^{**}) = .p^B$.

Notice that, due to imperfect competition on the issuers' side, a^{**} is higher than Baxter's interchange fee a_0:

$$f(c^B - a^{**}) > c^B - a^{**} \Rightarrow a^{**} > b^S - c^S = a_0.$$

This determines the socially optimum IF when $a^{**} < \bar{a}$, i.e. when a^* is acceptable to the sellers. When on the contrary the constraint that sellers accept the card is binding, the socially optimal IF is equal to $\bar{a}$. Thus:

Proposition 2 *The socially optimal level a^* of the interchange fee is equal to the minimum of a^{**} (which leads to a perfect internalization of the payment externality by the buyer) and $\bar{a}$ (the maximum level that is compatible with sellers' acceptance).*

A comparison of Propositions 1 and 2 shows that when sellers' willingness to pay for cards is low ($\bar{a}$ smaller than a^{**}), the privately optimal (Proposition 1) and socially optimal (Proposition 2) levels of the IF coincide: they are both equal to $\bar{a}$. On the other hand, if sellers' willingness to pay is high $\bar{a} > a^{**}$, the privately optimal IF $\bar{a}$ exceeds the socially optimal level a^*: there is overprovision of card payment services. When $\alpha = 0$ (no buyer is a priori informed about sellers' cards acceptance policy), we are back to Baxter's case: (5)

shows that $\bar{a} = a_0$, while (8) gives $a^{**} > a_0$. In this case, privately optimal and socially optimal IFs coincide, since $\bar{a} < a^{**}$.

4 The impact of costless surcharging

We have already observed that differentiated pricing of card and cash payments is seldom used in practice even when payment cards networks do not prohibit it, probably because of transaction costs. However, for the sake of the argument, we examine in this section the consequences of costless surcharging in the model of an imperfectly competitive banking industry developed in Section 3. Recall that in this model, sellers compete in prices and services offered to buyers. Costless surcharging implies that at equilibrium, sellers charge an additional amount of $p^S - b^S$ (the fee they pay to their bank minus their incremental benefit of accepting cards) whenever a buyer pays by card. The total price of card services for the buyer is then $\pi^B = p^B + p^S - b^S$. We denote by $D^B(\pi^B)$ the demand function[24] for card payments by buyers, that is, the proportion of buyers for which $b^B \geq \pi^B$:

$$D^B\left(\pi^B\right) \equiv \int_{\pi^B}^{+\infty} h^B\left(b^B\right) db^B .$$

When sellers can costlessly surcharge, they never benefit from refusing card payments. The total volume of card payments is then only determined by buyers' demand: it is equal to $D^B(p^B + p^S - b^S)$. With perfect surcharging and in the absence of transaction costs, volume only depends on the total price $(p^B + p^S)$ for card services and not anymore on the price structure. Similarly, the net margin of issuers $p^B - c^B + a$ only depends on total price and total cost. This is because $p^S = c^S + a$ (since acquirers are competitive), thus:

$$p^B - c^B + a = p^B + p^S - c.$$

The level of the IF ceases to play any role (neutrality). The total price of card services (which is here entirely borne by buyers) is then determined by the imperfect competition equilibrium between issuers. The demand function is $D^B(\pi^B)$ and the margin of issuers is $p^B + p^S - c = \pi^B - (c - b^S)$. Thus by definition of the function $f(\cdot)$, the total price paid by buyers at equilibrium is $f(c - b^S) > c - b^S$, implying under-provision of card services. The comparison with the case where surcharging is banned by the network (No Surcharge Rule, referred to as NSR hereafter) is thus ambiguous:

Proposition 3 *If surcharging is costless, banning the NSR may or may not increase welfare, depending on issuers' market power and sellers' resistance.*

This result is illustrated by the following figure, which represents social welfare W as a function of π^B:

$$W\left(\pi^B\right) = \int_{\pi^B}^{+\infty} \left(b^B + b^S - c\right) g\left(b^B\right) db^B .$$

[24] Throughout this article, demand functions are supposed to be log-concave, in order to ensure the quasi-concavity of profit functions.

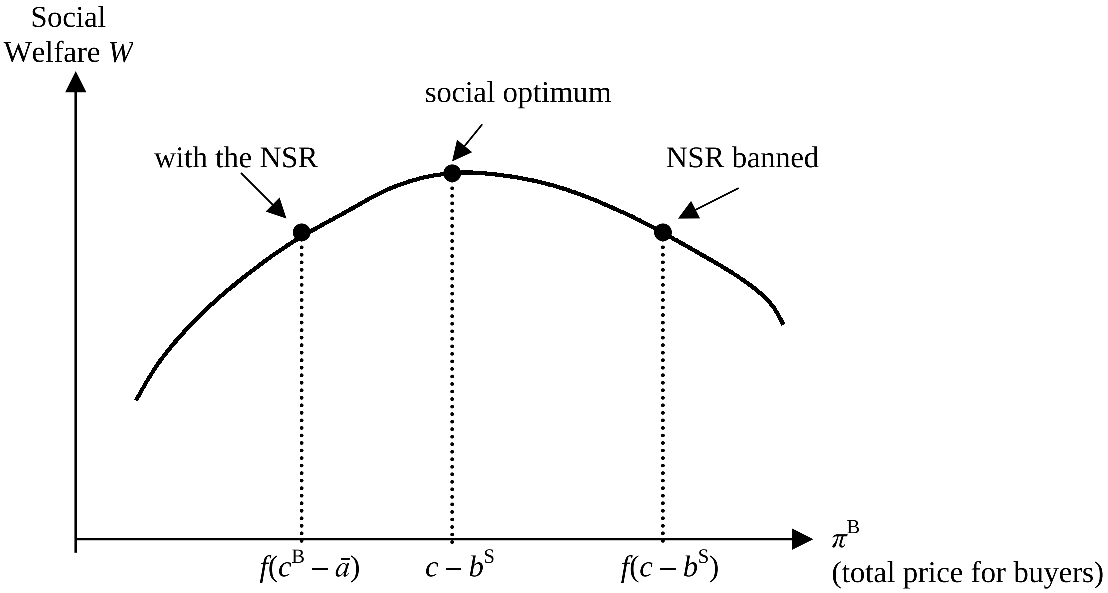

Figure 5: The (ambiguous) impact of banning the NSR when surcharging is costless (case where $\bar{a} > a^*$). Notice that $c - b^S = f(c^B - a^*)$

W is a concave function of π^B, with a maximum for $\pi^B = c - b^S$. Figure 5 illustrates the case where $f(c^B - \bar{a})$ (the price of card services when the NSR is in place) is lower than the optimal price $c - b^S = f(c^B - a^{**})$. This corresponds to the first configuration in Proposition 2 ($a^* = a^{**} < \bar{a}$), where the privately optimal IF is too high. Even in this case, banning the NSR does not necessarily increase welfare. This is because $f(c - b^S)$ (the price of card services when the NSR is in place) is higher than $c - b^S$. So both situations correspond to an inefficient usage of cards, but it is not possible to determine which one gives a higher value of social welfare W, without further information on the parameters of the model.

However, Figure 5 suggests the following intuitions (all else being given):

- When issuers' market power is large (i.e. $f(\gamma) - \gamma$ is large for all γ) then both $f(c^B - \bar{a})$ and $f(c - b^S)$ move to the right: banning the NSR is likely to be welfare decreasing.

- When sellers' resistance is strong (i.e. $\bar{a}$ is small) then $f(c^B - \bar{a})$ moves to the left: banning the NSR is likely to be welfare increasing.

Wright (2002a) extends the analysis of Rochet and Tirole (2002) by looking at alternative models of sellers' behavior. He shows that the impact of banning the NSR also depends on the market power of sellers. He looks first at the case where buyers buy a large number of goods and services, each sold by a monopoly provider. He assumes that buyers' preferences for these goods and services are identical: buyers all have the same valuation v for each of these goods or services. With the NSR in place (i.e. same price for cash and card payments), and provided that a sufficient fraction of buyers do not have cards[25] (or do not want to use them), monopolistic sellers charge the maximum price (i.e. $p = v$) that induces all potential consumers by buy. Then sellers accept cards if and only if $p^S = c^S + a \leq b^S$. Notice that since they are in a monopoly position, sellers are not afraid of losing customers if they refuse cards like in Baxter (1983), but contrary to what happens in Rochet and Tirole (2002). Since issuers' profit increases with the IF, the network selects the maximum level of IF that is compatible with sellers' acceptance, i.e. Baxter's IF:

$$a_0 = b^S - c^S.$$

The price set by issuers for card services is $p^B = f(c^B - a_0) = f(c - b^S)$, and there is underefficient usage of cards, due to issuers' market power.[26]

When the NSR is banned, the consequences differ from those in Rochet and Tirole (2002). Indeed if monopolistic sellers can costlessly surcharge, they pass through to buyers *more* than their net cost $p^S - b^S$. This entails an additional bias towards underprovision of card services, due to a double distortion on their pricing, a consequence of market power of both sellers and issuers. Notice that, also in this case, the interchange fee becomes neutral and ceases to play any part.[27]

More specifically, monopolistic merchants charge $p_{\text{cash}} = v$ for cash payments and $p_{\text{card}} = v + s_{\text{m}}$ for card payments, where s_{m} is the monopolistic surcharge, obtained by maximizing the profit realized by sellers on card payments:[28]

$$s_m = \arg\max_s \left(s - p^S + b^S\right) D^B\left(s + p^B\right).$$

The resulting total price of card services for buyers is $s_{\text{m}} + p^B = \pi_m^B$. Notice that

$$\pi_m^B = \arg\max_{\pi^B} \left(\pi^B - p^B - p^S + b^S\right) D^B\left(\pi^B\right).$$

π_m^B is a function of $p = p^B + p^S$ only and not on the price structure. More precisely, it is equal to $f_{\text{m}}(p - b^S)$, where $f_{\text{m}}(\gamma)$ realizes[29] $\max_{\text{f}}(f - \gamma) D^B(f)$. Like before, issuers' net margin $(p^B + a - c^B)$ is equal to $p - c$ and does not depend on the price structure, but only on the total price p. This explains the neutrality of IFs. However, the demand for card services by buyers is smaller than in Rochet-Tirole (2002), due to the monopoly power of sellers:

[25] Otherwise sellers may want to charge higher prices, that are acceptable only to cardholders.

[26] Interestingly, p^B coincides with the price selected by issuers in Rochet and Tirole (2002) when the NSR is banned.

[27] Gans and King (2003) have established that this neutrality of IFs (when costless surcharging is feasible) is a very general property.

[28] This analysis assumes away the existence of fixed fees for cardholders. If there is such a fee (however small) and if b^B is known *ex-ante* by cardholders, the marginal cardholder anticipates a negative surplus, which leads to a complete unravelling, in the spirit of Williamson's hold-up problem: consumers decide not to hold cards, anticipating they will end-up with a negative surplus. A more reasonable way to introduce fixed fees (or fixed costs) is to assume that b^B is not known *ex-ante* by the cardholder, in which case double marginalisation occurs.

[29] $f_{\text{m}}(\cdot)$ corresponds to the particular form taken by the function $f(\cdot)$ defined earlier, in the case of a monopolistic issuer. However here, it is the seller who behaves monopolistically.

$\Delta^B(p) \equiv D^B(\pi_m^B) = D^B(f_m(p-b^S))$.

The situation can be depicted by a chain of imperfectly competitive markets, as illustrated by figure 6.

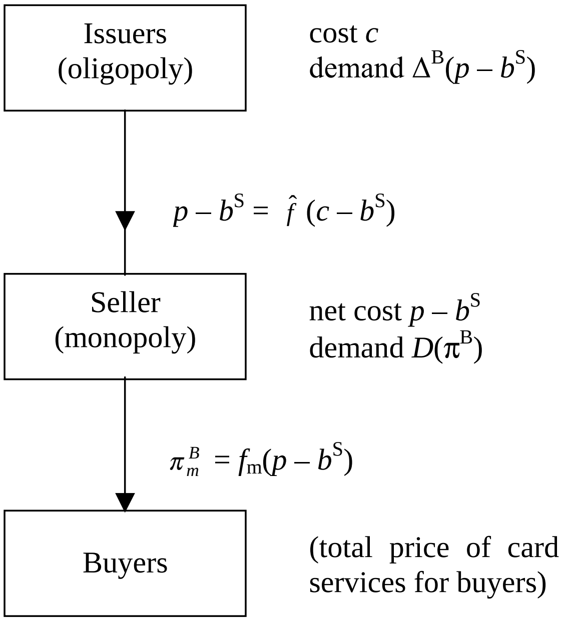

Figure 6: Excessive surcharging by a monopolistic seller

The resulting total price of card payments for buyers is denoted $\hat{f}(c-b^S)$. It is higher than $f(c - b^S)$, because of double distortions.[30] These results are illustrated in Figure 7.

Figure 7 is a variant of Figure 5: it also represents social welfare as a function of π^B, and the impact of banning the NSR but here sellers have monopoly power. Figure 7 shows that, in this case (and contrary to the ambiguous situation depicted in Figure 5, where sellers are Hotelling competitors), social welfare unambiguously decreases if the NSR is banned. Indeed, with monopolistic sellers, the NSR prevents excessive surcharging and increases card payment volume and social welfare. Notice also that the socially optimal and privately optimal IF coincide with a_0 (since $c - b^S = c^B - a_0$), the maximum level that is compatible with sellers' acceptance. Due to the greater market power of sellers, both issuers and the regulator would in this case select the maximum possible IF.

Wright (2002a) also considers the other extreme case of perfect (undifferentiated Bertrand) competition between sellers, again in a frictionless context in which sellers don't face costs of having multiple prices and buyers are perfectly informed of who takes the

[30] This situation (monopolistic seller and oligopolistic issuers) is a variant of the chain of monopolies analyzed in the Industrial Organization literature on intermediaries (see for example Tirole, 1988). There are two reasons why $\hat{f}(c-b^S) > f(c-b^S)$: the mark-up extracted by the monopolistic seller at the bottom of the chain, and the fact that Δ^B is more elastic than D^B, which results from the log-concavity of D^B and the fact that $f_m(p - b^S) > (p - b^S)$.

card. In this case it is easy to see that sellers specialize: low price sellers (who charge only the marginal cost of production of the good) only accept cash payments whereas high price sellers (who charge also the net cost of cards $p^S - b^S$) only serve cardholders.[31] The NSR becomes ineffective and the level of the IF is again neutral.

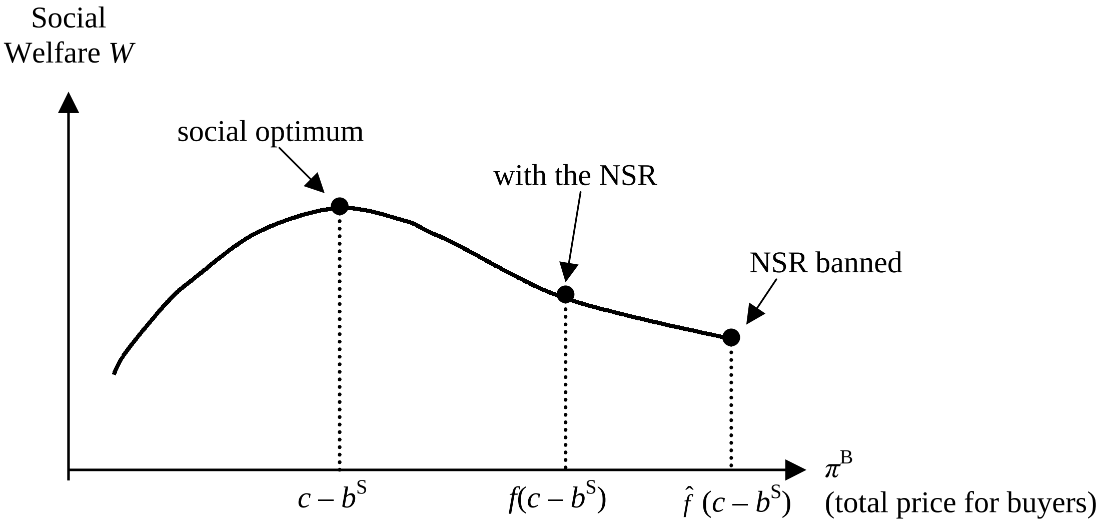

Figure 7: The impact of banning the NSR with monopolistic sellers

Let us summarize the results of this section:

- When costless surcharging is feasible, the level of the IF becomes neutral, and there is underprovision of card services (Rochet and Tirole, 2002, and Gans and King, 2003).
- If sellers are monopolistic (Wright, 2002a), the NSR partially corrects this underprovision: transaction volume and social welfare are increased.
- If sellers are Hotelling competitors (Rochet and Tirole, 2002), the NSR also leads to an increase in card transactions but the impact on social welfare is ambiguous.
- If sellers are perfect competitors (Wright, 2002a) the NSR has no impact on transaction volumes nor social welfare.

These results are represented in the following table:

[31] This is confirmed by casual empiricism: deep discount supermarkets often refuse cards, while regular supermarkets tend to accept them. I thank Yossi Spiegel for this remark.

Type of competition on downstream markets (sellers)	Impact of the NSR	Volumes	Welfare	Reference
monopolistic	limits inefficient surcharges	+	+	Wright (2002a)
Hotelling	decreases cardholder fees	+	ambiguous	Rochet and Tirole (2002)
Bertrand	neutral	=	=	Wright (2002a)

Figure 8: The impact of the NSR according to the type of competition in downstream markets

5 System competition

We have seen in Section 3 that the interchange fee $\bar{a}$ selected by a monopoly association was not necessarily different from the socially optimum IF a^*. Proposition 2 shows that when acquirers are perfectly competitive and merchants homogenous, the privately optimal IF $\bar{a}$ is also socially optimal if the market power of issuers is sufficiently important. However when the downstream market for buyers is sufficiently competitive, the IF $\bar{a}$ chosen by a monopoly association is too high (and buyers' prices are too low), as compared with the social optimum. The objective of this section is to see how this situation is modified by system competition.

Rochet and Tirole (2002) consider the case of two competing associations (i = 1,2), offering perfectly substitutable payment card services. The only difference to the timing above is in the first stage, where the two associations simultaneously choose their interchange fees a_1 and a_2. For simplicity, they assume that downstream competition between issuers and between acquirers is not affected by system competition. Also, they neglect duality issues,[32] i.e. the fact that some issuers and/or some acquirers may be members of the two associations.

The interchange fees selected by the associations correspond to the perfect Nash equilibrium of the game associated to this new timing. Rochet and Tirole (2002) do not solve this game but just notice that when each buyer only has a single card, IFs are not affected by system competition: $a_1 = a_2 = \bar{a}$ is the equilibrium of the game between associations. This is because the incentives of the associations and of the buyers are perfectly aligned: both want to maximize the IF under the constraint that the card is accepted by the sellers.

The situation changes when (at least some) buyers have two cards: $a_1 = a_2 = \bar{a}$ is no more an equilibrium. Indeed if system 1 chooses a slightly lower IF, the sellers (who were indifferent between accepting cards or not) reject card 2, since (at least with some probability) buyers have card 1 in their wallet, and sellers get a positive surplus from card 2 transactions.

[32] On these issues, see Hausman et al. (2003).

Guthrie and Wright (2003) compute the Nash equilibrium IFs in the case when $\alpha = 1$ and buyers hold both cards. They find that both networks select the interchange fee a^c (the "competitive" interchange fee) that maximizes total users' surplus. As a function of p^B and p^S, this total users' surplus writes:

$$\phi = \int_{p^B}^{+\infty} \left(b^B + b^S - p^B - p^S\right) h\left(b^B\right) db^B .$$

Users' prices p^B and p^S are determined by the level a of the IF, and by the intensity of downstream competition. To simplify the analysis suppose again that acquirers are competitive:

$p^S = c^S + a$,

and that issuers' margins are constant:

(9) $f(\gamma) \equiv \gamma + m$.

This implies that:

$p^B \equiv c^B - a + m$,

so that total users' price is independent of a:

$p^B + p^S = c + m$.

In this case total users' surplus has a simple expression as a function of the IF a:

$$\phi(a) = \int_{c^B - a + m}^{+\infty} \left(b^B + b^S - c - m\right) h\left(b^B\right) db^B .$$

The competitive interchange fee realizes the maximum of this function:

$$\dot{\phi}\left(a^c\right) = \left(c^B - a^c + m + b^S - c - m\right) h\left(c^B - a^c + m\right) = 0 .$$

After simplifications we get:

(10) $a^c = b^S - c^S$.

Proposition 4 *When total user' price is independent of the IF, and issuers are imperfectly competitive, the competitive interchange fee a^c is lower than the socially optimal IF a^*.*

Thus two sided markets have rather peculiar properties: in the particular case considered here (more market power on the issuers' side than on the acquirer's side), competitive IFs are too low while monopoly IFs are either too high or socially optimal.

These results are illustrated by the following figure:

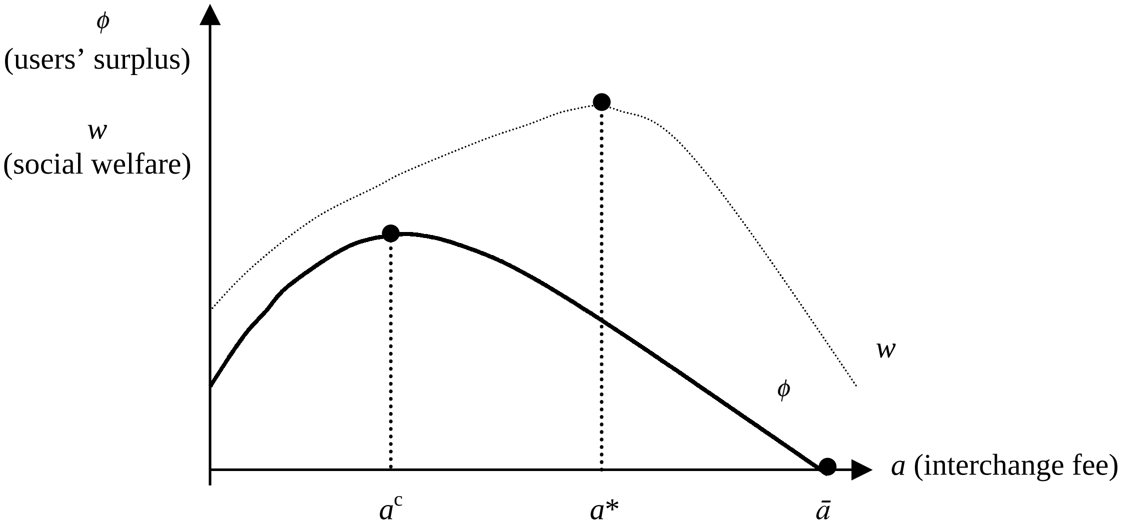

Figure 9: Comparison of different levels of the interchange fee

Notes: a^c: competitive, a*: socially optimal, ā: monopoly

6 Balancing the two sides of the market

So far, we have considered the case where sellers are homogenous: they all have the same incremental benefit b^S for a card versus a cash payment, while buyers have different valuations. This introduces an asymmetry between buyers and sellers. In particular, in this context the card network typically sets the IF at the highest level that sellers can accept. In practice, the situation is likely to be more balanced. A more symmetric model is proposed by Schmalensee (2002), who assumes that transaction volumes on the card network are the product of two "quasi-demand" functions. In our notation:

(11) $Q = D^B(p^B)D^S(p^S)$.

This specification is consistent with our previous model if we assume that b^S varies across sellers, and is distributed according to a density $g(b^S)$. In this case, the proportion of sellers with $b^S \geq p^S$ is

$$D^S(p^S) = \int_{p^S}^{+\infty} g(b^S)db^S .$$

Then (11) means that a card transaction takes place if and only if the buyer wants to use his card ($b^B \geq p^B$), and the seller accepts the card ($b^S \geq p^S$).[33] In any other case,

[33] For the moment, we adopt Baxter's assumption that sellers' acceptance decision is only influenced by direct costs and benefits (equivalently, $\alpha = 0$).

payment is by cash. It is also assumed that the buyer and the seller are independently drawn from their respective populations, which leads to the multiplicative specification (11). Notice finally that, in contrast with search models à la Baye and Morgan (2001) or Caillaud and Jullien (2001), we take as given the matching process between buyers and sellers, and focus on the proportion Q of trades that is settled by a card payment.

The (finite) elasticities of card transactions volume Q with respect to both prices modify the determination of optimal IFs, both from a private and a social point of view. Schmalensee (2002) considers for example the case of a network with only two banks: a monopoly issuer and a monopoly acquirer, who bargain on the level of the IF. Once this IF has been chosen, each bank chooses the final price on its side of the market so as to maximize its profit. The fact that p^B and p^S are not jointly determined creates a source of inefficiency,[34] analogous to a problem of moral hazard in teams. The purpose of the IF is not to correct this inefficiency,[35] but to "balance" demands on both sides of the market. Even if all the bargaining power within the card network is on one side of the market (say, the issuer's side) the IF will not necessarily be set at an excessive level, as compared with the socially optimal IF.

Rochet and Tirole (2003) extend Schmalensee's analysis by looking at price determination in general two sided markets. Their results can be applied to the payment card context, in the case where issuers' margins are constant and sellers are not strategic (equivalently, $\alpha = 0$, consistently with Baxter's assumptions). The socially optimal IF is obtained by maximization of social welfare:

$$(12)\quad W\left(b_m^B, b_m^S\right) = \int_{b_m^B}^{+\infty} \int_{b_m^S}^{+\infty} \left(b^B + b^S - c\right) g\left(b^S\right) h\left(b^B\right) db^S\, db^B ,$$

where b_m^B (respectively b_m^S) denotes the marginal buyer (respectively seller), and, due to our assumptions that issuers' margins are constant:

$$(13)\quad b_m^B = p^B = c^B - a + m ,$$

and acquirers are competitive:[36]

$$(14)\quad b_m^S = p^S = c^S + a .$$

The socially optimal interchange fee is thus characterized by the first order condition: $\dfrac{\partial W}{\partial b_m^B} = \dfrac{\partial W}{\partial b_m^S}$.

Using (12) we obtain:

[34] Indeed, p^B and p^S separately maximize the profit of each bank, without internalizing the impact on the bank on the other side of the market. As a result, total price is lower than the price that would maximize the aggregate profit of the network.

[35] This is because in the model, the IF only affects relative prices, not total price.

[36] The analysis could easily be extended to the case where acquirers also have positive (but constant) margins. The crucial simplifying assumption is that the IF does not affect total price $p^B + p^S$.

$$\frac{\partial W}{\partial b_m^B} = -\int_{b_m^S}^{+\infty} \left(b_m^B + b^S - c\right) g\left(b^S\right) h\left(b_m^B\right) db^S$$

$$= -\left(b_m^B + b_m^S + v^S\left(b_m^S\right) - c\right) h\left(b_m^B\right) D^S\left(b_m^S\right)$$

since $v^S\left(b_m^S\right) = E\left(b^S \middle| b^S \geq b_m^S\right) - b_m^S$.

Finally, using $-h\left(b_m^B\right) = \dot{D}^B\left(b_m^B\right)$ and $b_m^B + b_m^S = c$, we obtain:

$$\frac{\partial W}{\partial b_m^B} = v^S\left(b_m^S\right) \dot{D}^B\left(b_m^B\right) D^S\left(b_m^S\right)$$

By symmetry, we also have:

$$\frac{\partial W}{\partial b_m^S} = v^B\left(b_m^B\right) \dot{D}^S\left(b_m^S\right) D^B\left(b_m^B\right).$$

Thus, the socially optimal interchange fee is characterized by the condition:

$$v^S \dot{D}^B D^S = v^B \dot{D}^S D^B,$$

which is equivalent to:

$$(15) \quad \frac{\eta^B}{p^B v^B} = \frac{\eta^S}{p^S v^S},$$

where $\eta^k = -\frac{p^k \dot{D}^k}{D^k} \left(k = B, S\right)$ denotes the elasticity of quasi-demands and $v^k = E[b^k - p^k | b^k \geq p^k]$ ($k = B,S$) denotes the average net surplus of users on side k of the market. This condition can also be written in terms of relative prices:

$$\frac{p^B}{p^S} = \left(\frac{\eta^B}{\eta^S}\right) \Big/ \left(\frac{v^B}{v^S}\right).$$

Proposition 5 *When total users' price is independent of the IF and sellers are not strategic (or buyers not informed about card acceptance, i.e. $\alpha = 0$) the socially optimal interchange fee is obtained when relative prices are equal to the ratio of demand elasticities divided by the ratio of average net surpluses on each side of the market.*

Equation (15) is not specific to payment card networks. It constitutes a general pattern of two sided markets. In such markets, the socially optimal allocation of costs[37] across the

[37] This is only a second best (Ramsey) optimum since budget balance is required.

two sides of the market is influenced by two considerations: relative demand elasticities (marginal users) and relative net surpluses (average users).

Rochet and Tirole (2003) also study price determination in several other contexts (like competition between proprietary networks or open networks) and show that privately determined IFs are not systematically biased in one direction or the other. For example if we stick to the limit case of Proposition 5 (constant total price, non strategic sellers), the privately optimal IF $\bar{a}$ is obtained by maximizing card transaction volume

$$Q(a)=D^B(c^B - a + m)D^S(c^S + a).$$

Assuming that quasi-demands are log concave (so that Q is itself log concave), $\bar{a}$ is then characterized by

$$-\dot{D}^B D^S + D^B \dot{D}^S = 0\ ,$$

or

$$(16)\quad \frac{\eta^B}{p^B} = \frac{\eta^S}{p^S}\ .$$

Clearly, this differs from condition (15), which characterizes the socially optimal IF a^*. However, there is no systematic bias. In fact, the situations where $\bar{a} > a^*$ (the IF chosen by the network is too high) can be easily characterized. Indeed, they correspond to the cases where the sellers' price p^S determined by (15) is higher than that determined by (16). Since $\frac{p^k}{\eta^k} = -\frac{D^k}{\dot{D}^k}$ is a decreasing function of p^k (this is because log D^k is concave) for $k = B,S$, this is equivalent to $v^S < v^B$, hence the result:

Proposition 6 *When total price is constant and sellers are not strategic (or $\alpha = 0$), the IF $\bar{a}$ chosen by the network is higher than the socially optimal IF a^* (which implies that buyers pay too little and sellers pay too much) if and only if, at $\bar{a}$ the average net surplus of buyers exceeds the average net surplus of sellers: $v^S < v^B$.*

Consider for example the case of log-linear demands:[38]

$$D^k\left(p^k\right) = \left(A^k - p^k\right)^{\varepsilon^k}, \quad k = B, S\ ,$$

where A^k and ε^k are positive constants. The privately optimal IF must satisfy the condition:

$$\frac{\eta^B}{p^B} = \frac{\eta^S}{p^S}\ .$$

This gives here:[39]

[38] We cannot take the example of demands with constant elasticities ($D(p) = Ap^{-\varepsilon}$) since they are not log concave. Notice that log concavity implies that demand elasticity increases with the price.

$$(17)\ \frac{\varepsilon^B}{A^B - p^B} = \frac{\varepsilon^S}{A^S - p^S}\ .$$

We have to compare the associated levels of average surpluses of buyers and sellers:

$$v^B = \frac{A^B - p^B}{1+\varepsilon^B} = \frac{\varepsilon^B}{(1+\varepsilon^B)}\left(\frac{\eta^B}{p^B}\right)$$

$$v^S = \frac{A^S - p^S}{1+\varepsilon^S} = \frac{\varepsilon^S}{(1+\varepsilon^S)}\left(\frac{\eta^S}{p^S}\right).$$

At the privately optimal IF $\bar{a}$, we have that $v^S < v^B$ if and only if

$$\frac{\varepsilon^S}{1+\varepsilon^S} < \frac{\varepsilon^B}{1+\varepsilon^B}\ ,$$

or $\varepsilon^S < \varepsilon^B$.

Proposition (6) shows that $\bar{a} > a^*$ if and only if $v^S < v^B$, which is thus equivalent to $\varepsilon^S < \varepsilon^B$. In particular, for linear demands ($\varepsilon^S = \varepsilon^B = 1$) the privately optimal and socially optimal IF coincide. On the other hand, if sellers tend to be homogenous ($\varepsilon^S \to 0$), privately optimal IFs are likely to be too high.

Wright (2001, 2002b) incorporates the multiplicative demand specification of Schmalensee (2002) into the model of Rochet and Tirole (2002). He takes into account the strategic behavior of sellers (and assumes that $\alpha > 0$). Let us summarize his contributions. For simplicity we stick to the case where issuers' margins are constant and acquirers are competitive.

Equations (13) and (14) remain valid:

$$(18)\ b_m^B = p^B = c^B - a + m\,,$$

$$(19)\ p^S = c^S + a,$$

except that, with strategic sellers, the marginal seller is given by:

[39] Since $p^B + p^S = c$, condition (17) gives the prices for buyers and sellers:

$$p^B = A^B - \frac{\varepsilon^B}{\varepsilon^B + \varepsilon^S}\left(A^B + A^S - c\right)$$

$$p^S = A^S - \frac{\varepsilon^S}{\varepsilon^B + \varepsilon^S}\left(A^B + A^S - c\right)$$

The privately optimal IF is thus:

$$\bar{a} = p^B - c^B = \frac{\varepsilon^S\left(A^B - c^B\right) - \varepsilon^B\left(A^S - c^S\right)}{\varepsilon^S + \varepsilon^B}.$$

(20) $b_m^S = p^S - \alpha v^B \left(p^B\right).$

As usual, Wright assumes that the network sets the IF *a* so as to maximize the total profit of its members, here equal to:

(21) $B(a) = mD^B\left(b_m^B\right)D^S\left(b_m^S\right).$

The characterization of Proposition 6 remains valid in this context:

Proposition 7 *When total users' price is constant and sellers are strategic* $\alpha > 0$*) the IF ā chosen by the network is higher than the socially optimal IF* a^* *if and only if, at ā, the average net surplus of buyers exceeds the average net surplus of sellers* $v^S < v^B$*. When demands are log-linear, ā is always higher than* a^**.*

Proposition 7 shows that the first model considered in Rochet and Tirole (2002) is somewhat biased towards excessive IFs, since it assumes $v^S = 0 < v^B$ (homogenous sellers, but heterogenous buyers). The more balanced specifications of Schmalensee (2002), Wright (2001), (2002b) and Rochet and Tirole (2003) lead instead to configurations where IFs can be too low ($\bar{a} < a^*$).

For example with competitive downstream markets and heterogenous sellers, Wright (2001) shows that privately optimal IFs are too high (from a social welfare viewpoint) whenever the average transactional benefits of the sellers who accept cards ($v^S + b_m^S$ in our notation) is less than the sellers price p^S. Since $b_m^S = p^S - b^B$, this is equivalent to our condition $v^S < v^B$.

However, the fact that merchants also accept cards in order to attract more customers pushes towards high IFs:

Proposition 8 *The greater the competitive edge obtained by sellers who accept cards (measured by* α*) the more likely the card network sets an interchange fee at a higher level than the social optimum.*

Finally, let us mention the interesting (and surprising) result of Wright (2000b) who shows that the first best optimum (i.e. without imposing budget balance) can be implemented when $\alpha = 1$, by an appropriate choice of the interchange fee. Indeed, the conditions for social welfare (unconstrained) maximization are:

(22) $\dfrac{\partial W}{\partial b_m^B} = \left[b_m^B + b_m^S + v^S\left(b_m^S\right) - c\right]\dot{D}^B D^S = 0$,

and

(23) $\dfrac{\partial W}{\partial b_m^S} = \left[b_m^B + b_m^S + v^B\left(b_m^B\right) - c\right]\dot{D}^S D^B = 0$,

This is equivalent to:

(24) $v^S\left(b_m^S\right) = v^B\left(b_m^B\right) = c - b_m^B - b_m^S$.

Equation (24) shows well that optimality of cards usage requires that marginal users be subsidized (b_m^B + b_m^S < c). We now show that this is compatible with budget balance of the network (p^B + p^S = c) if sellers are strategic and buyers are perfectly informed about card acceptance (α = 1). Indeed, when banks (issuers and acquirers) are perfectly competitive, we have seen that

(25) $b_m^B = p^B = c^B - a$,

and

(26) $b_m^S = p^S - \alpha v^B\left(b_m^B\right) = c^S + a - \alpha v^B\left(b_m^B\right)$.

As noticed by Wright (2002b), this is compatible with (24) if and only if α = 1. Indeed, by adding (25) and (26) we obtain:

$$b_m^B + b_m^S = c - \alpha v^B\left(b_m^B\right),$$

which implies (23) if and only if α = 1. It is then possible to select the IF a in such a way that

$$v^B(c^B - a) = v^S(c^S + a - v^B(c^B - a)),$$

so that the efficiency condition in (24) is fully satisfied.

This has to be contrasted with the case α = 0, already considered in Proposition 4. In this case W was maximized under the constraint

$$b_m^B + b_m^S = c,$$

obtained by adding (13) and (14), and equivalent to budget breaking at the network level. The conditions characterizing the (second best) optimal IF were:

$$\frac{p^B}{\eta^B} v^B\left(p^B\right) = \frac{p^S}{\eta^S} v^S\left(p^S\right),$$

with $p^B = c^B - a$, and $p^S = c^S + a$.

By contrast, when α = 1, sellers accept to pay more than their marginal benefit. For strategic reasons, they internalize the average benefit of cardholders. The net cost they incur is then transferred to their customers (including those paying by cash) in the form of a price increase. This mechanism serves as a “budget breaker” since b_m^B + b_m^S is now less than c, and it is possible to select the IF in such a way that the unconstrained maximum of social welfare W is attained.

It is interesting to recall that several commentators have criticized the fact that, due to the NSR, cash users bear some fraction of the cost of card transactions.[40] Wright's result sheds light on an important phenomenon that had been previously overlooked: The NSR allows for an internalization by sellers of the positive surplus they generate on the buyers' side when they accept cards. However this internalization is not perfect because of the market power of issuers. As we saw in Section 5 (in particular Figure 9), competitive card networks select IFs that maximize total users' surplus, and therefore are lower than social optimal IFs.

7 Directions for future research

We now possess a relatively robust and tractable theoretical model of the payment card industry.[41] This model has already been used by Rochet and Tirole (2003) and Armstrong (2002) to study other examples of two-sided markets. However, more research on the payment card industry itself is needed, in particular along the following directions:

- Costly Surcharging: so far, the analysis has focused on two polar cases. Either surcharges are impossible (perfect two sided market), or are costless (neutrality). It would be important to extend the analysis of the determinants of privately optimal and socially optimal IFs to a case where surcharges are allowed but costly, so that they only a fraction of sellers charge differentiated prices for cash and card payment.

- Competition between several payment instruments: for tractability, we have considered cash as the only alternative to card payments. Other possibilities include checks, electronic money and different forms of cards. Rochet and Tirole (2003) have characterized equilibria when two symmetric payment card networks compete. An extension to asymmetric equilibria, say between debit and credit cards, or other payment instruments would be useful.

More generally, the models presented in this article shed light on the likely impact of the regulations of payment card networks that are currently envisaged in many countries. They are also a good starting point for analyzing the on-going antitrust cases in the U.S. and elsewhere.

[40] It is true that cash users are often less wealthy than card users, and therefore may be hurt more by a uniform price increase. It is also true that we have assumed inelastic demands for final goods and services: with elastic demands, including the cost of card payments into the final price of goods and services has a negative impact on social welfare. However, this negative impact is likely to be counterbalanced by an increase in demand coming from cardholders.

[41] We have left aside two important aspects of payment cards, which are examined in separate branches of the literature: credit functionalities (see e.g. Ausubel 1991; Chakravorti and Emmons 2001; Chakravorti and To, 2000); the possibility to withdraw cash in ATM networks (see e.g. Baker, 1995; Kim, 1998; Matutes and Padilla 1998; McAndrews, 1996; McAndrews and Rob, 1996; Donze and Dubec, 2002).

8 References

Ausubel, L.M. (1991) "The Failure of Competition in the Credit Market," *American Economic Review*, 81: 50-81.

Armstrong, M. (2002) "Competition in Two-Sided Markets," presentation at ESEM meeting, Venice, August.

Baker, D.I. (1995) "Shared ATM Networks. The Antitrust Dimension," *Federal Reserve Bank of St. Louis Review*, 77: 5-17.

Balto, D.A. (2000) "The Problem of Interchange Fees: Costs without Benefits?" *European Competition Law Review*, 21: 215-224.

Baxter, W.F. (1983) "Bank Interchange of Transactional Paper: Legal Perspectives," *Journal of Law and Economics*, 26: 541-588.

Baye, M., and J. Morgan (2001) "Information Gatekeepers on the Internet and the Competitiveness of Homogenous Product Markets," *American Economic Review*, 91: 454-474.

Caillaud, B. and B. Jullien (2003) "Chicken and Egg: Competition among Intermediation Service Providers," *RAND Journal of Economics*, forthcoming.

Carlton, D.W. and A.S. Frankel (1995) "The Antitrust Economics of Payment Card Networks," *Antitrust Law Journal*, 63: 643-668.

Chakravorti, S. and W.R. Emmons (2001) "Who Pays for Credit Cards?" Federal Reserve Bank of Chicago Emerging Payments Occasional Paper Series, EPS-2001-1.

Chakravorti, S. and A. Shah (2001) "A Study of the Interrelated Bilateral Transactions in Credit Card Networks," Federal Reserve Bank of Chicago Emerging Payments Occasional Paper Series, EPS-2001-2.

Chakravorti, S. and T. To (2000) "Towards a Theory of Merchant Credit Card Acceptance," mimeo, Federal Reserve Bank of Chicago.

Chang, H.H. and S. Evans (2000) "The Competitive Effects of the Collective Setting of Interchange Fees by Payment Card Systems," *The Antitrust Bulletin*, Fall, 641-677.

Donze, J. and I. Dubec (2002) "Access Pricing in Shared ATM Networks," discussion paper, IDEI Center for Payment Cards.

Evans, D. S. and R.L. Schmalensee (1993) *The Economics of the Payment Card Industry*. National Economic Research Associates.

Evans, D. S. and R.L. Schmalensee (1995) "Economic Aspects of Payment Card Systems and Antitrust Policy Toward Joint Ventures," *Antitrust Journal Law*, 63: 861-901.

Evans, D. S. and R.L. Schmalensee (1999) *Paying with Plastic: The Digital Revolution in Paying and Borrowing*. Cambridge, MIT Press.

Evans, D.S. (2002) “The Antitrust Economics of Two-Sided Markets,” mimeo, AEI-Brookings Joint Center.

Frankel, A.S. (1998) “Monopoly and Competition in the Supply and Exchange of Money,” *Antitrust Law Journal*, 66: 313-361.

Gans, J.S. and S.P. King (2001) “Regulating Interchange Fees in Payment Systems,” mimeo, University of Melbourne.

Gans, J.S. and S.P. King (2002) “A Theoretical Analysis of Credit Card Regulation,” mimeo, University of Melbourne.

Gans, J. and S.P. King (2003) “The Neutrality of Interchange Fees in Payment Systems,” *Topics in Economic Analysis and Policy*, Vol 3, Article 1.

Guthrie, G. and J. Wright (2003) “Competing Payments Schemes,” mimeo, University of Auckland.

Hausman, J., Lenonard, G. and J. Tirole (2003) “On Non-exclusive Membership in Competing Joint Ventures,” *RAND Journal of Economics*, forthcoming.

Katz, M. and C. Shapiro (1985) “Network Externalities, Competition, and Compatibility,” *American Economic Review*, 75: 424-440.

Katz, M. (1986) “Technology Adoption in the Presence of Network Externalities,” *Journal of Political Economy*, 94: 822-841.

Kim, J. (1998) “The Impact of Proprietary Positions and Equity Interest in the Pricing of Network ATM Services,” mimeo, MIT.

Matutes, C. and A.J. Padilla (1994) “Shared ATM Networks and Banking Competition,” *European Economic Review*, 38: 1113-1138.

McAndrews, J.J. (1996) “Retail Pricing of ATM Network Services,” Working Paper 96-12, Federal Reserve Bank of Philadelphia.

McAndrews, J.J. and R. Rob (1996) “Shared Ownership and Pricing in a Network Switch,” *International Journal of Industrial Organization*, 14: 727-745.

Rochet, J.C. and J. Tirole (2002) “Cooperation among Competitors: Some Economics of Payment Card Associations,” *RAND Journal of Economics*, 33: 1-22.

Rochet, J.C. and J. Tirole (2003) “Platform Competition in Two-Sided Markets,” *Journal of the European Economic Association*, forthcoming.

Schmalensee, R. (2002a) “Payment Systems and Interchange Fees,” *Journal of Industrial Economics*, 50: 103-122.

Schmalensee, R. (2002b) “Interchange Fees: A Review of the Literature,” mimeo, MIT.

Schwartz, M. and D. Vincent (2000) “The No-Surcharge Rule in Electronic Payments Markets: A Mitigation of Pricing Distortions,” mimeo, Georgetown University.

Schwartz, M. and D. Vincent (2001) “Same Price, Cash or Card: Vertical Control by Payment Networks,” mimeo, Georgetown University.

Small, J. and J. Wright (2000) “Decentralized Interchange Fees in Open Payment Networks: An Economic Analysis,” mimeo, University of Auckland.

Tirole, J. (1988) “The Theory of Industrial Organization,” Cambridge, MIT Press.

Wright, J. (2000) “An Economic Analysis of a Card Payment Network,” mimeo, University of Auckland.

Wright, J. (2001) “The Determinants of Optimal Interchange Fees in Payment Systems,” Working Paper No. 220, Department of Economics, University of Auckland.

Wright, J. (2002a) “Optimal Card Payment Systems,” *European Economic Review*, forthcoming.

Wright, J. (2002b) “Pricing in Debt and Credit Card Schemes,” *Economics Letters*, forthcoming.

Wright, J. (2003) “Why do Firms Accept Credit Cards?” mimeo, University of Auckland.

9 Appendix: Mathematical proofs

Proof of Lemma 1: The left-hand side of inequality (5) is an increasing function of a. Indeed $\dot{v}^B$, the derivative of the net surplus is larger than –1,[42] while $\dot{f} \geq 1$ by assumption. Therefore the derivative of the left-hand side of inequality (5) with respect to a is at least equal to 1 – α and thus positive (see Figure 4). Therefore inequality (5) is satisfied, i.e. cards are accepted, if and only if $a \leq \bar{a}$, where $\bar{a}$ is the unique value of a that satisfies equality in (5). ■

Proof of Proposition 4: First notice that $\bar{a}$, the monopoly IF is defined implicitly by $\phi(\bar{a}) = 0$. This is because

$$\frac{\phi(a)}{D^B\left(p^B\right)} = v^B\left(p^B\right) + p^B + b^S - c - m \,.$$

Since $p^B + p^S = c + m$, this is zero when $p^S = b^S + b^B(p^B)$, or $a - v^B(f(c^B - a)) = b^S - c^S$, which coincides with the implicit definition of $\bar{a}$ (see Proposition 1) when $\alpha = 1$. Then recall (see Proposition 2) that the socially optimal interchange fee is the minimum of $\bar{a}$ and

[42] This is because $v^B + p^B = E[b^B | b^B \geq p^B]$ is increasing in p^B.

a^{**}, the value that maximizes social welfare. Given our specification of f – see (9) – we have an explicit formula for a^{**}

$$f(c^{\mathrm{B}} - a^{**}) = c - b^{\mathrm{S}} \Rightarrow a^{**} = b^{\mathrm{S}} - c^{\mathrm{S}} + m.$$

It is immediate that both $\bar{a}$ and a^{**} are greater than a^{c}. Thus $a^{*} = \min(\bar{a}, a^{**}) > a^{\mathrm{c}}$. ■

Proof of Proposition 7: The privately optimal IF is determined by the condition:

$$(27) \quad \frac{p^B}{\eta^B}\left(1+\alpha\dot{v}^B\right)=\frac{p^S+\alpha v^B}{\eta^S}.$$

On the other hand the condition for a socially optimal IF is

$$\frac{v^B p^B}{\eta^B}\left(1+\alpha\dot{v}^B\right)=\frac{v^S\left(p^S+\alpha v^B\right)}{\eta^S}.$$

Again, the condition for the privately optimal IF to be too high from a social viewpoint is $v^{\mathrm{B}} > v^{\mathrm{S}}$. ■

Proof of Proposition 8: With log concave demands, $p^{\mathrm{B}}/\eta^{\mathrm{B}}$ and $p^{\mathrm{S}}/\eta^{\mathrm{S}}$ are decreasing (respectively in p^{B} and p^{S}), and $\dot{v}^B < 0$. This implies that the price p^{S} determined by (27) (and the condition $p^{\mathrm{B}} + p^{\mathrm{S}} = c + m$) increases with α (and symmetrically that p^{B} decreases with α). This is intuitive: the greater α, the greater the sellers' willingness to pay for card services and the higher the ratio $p^{\mathrm{S}}/p^{\mathrm{B}}$.

Now excessively high IFs occur when $v^{\mathrm{S}} < v^{\mathrm{B}}$. Log concavity of demands implies that both $\dot{v}^S$ and $\dot{v}^B$ are negative. Since $dp^{\mathrm{B}}/d\alpha < 0$ and $dp^{\mathrm{S}}/d\alpha > 0$, we see that the function $v^{\mathrm{B}}[p^{\mathrm{B}}(\alpha)] - v^{\mathrm{S}}[p^{\mathrm{S}}(\alpha)]$ increases in α. In other words, the greater α, the weaker sellers' resistance and the more likely it is that the IF set by the network is excessively high. ■