The No Surcharge Rule and Card User Rebates:

Vertical Control by a Payment Network

by

Marius Schwartz

Daniel R. Vincent*

March 15, 2004

Abstract: The No Surcharge Rule (NSR) prohibits merchants from charging higher prices to consumers who pay by card instead of other means ('cash'). We analyze the NSR's effects when a card company faces local monopolist merchants. The NSR raises card company profit and harms cash users and merchants, while overall welfare decreases if and only if the ratio of cash to card users is sufficiently small. If the card company cannot grant rebates to card users, the NSR reduces total consumer surplus (of cash plus card users) and, if the cash market is sufficiently small, even card users lose. When rebates are feasible, the card company will grant them and raise its merchant fee, benefitting total consumer surplus and overall welfare but harming cash users compared to no rebates. An increase in the merchant's gross benefit from card rather than cash sales worsens the NSR's effects on overall welfare and total consumer surplus if card user rebates are not feasible, but the reverse holds if rebates are feasible.

JEL Classification: D42, G28, L42, L8.

---

* Schwartz: Department of Economics, Georgetown University, Washington DC 20057-1036, <schwarm2@georgetown.edu> Vincent: Department of Economics, University of Maryland, College Park, MD 20742 <dvincent@wam.umd.edu>. For helpful comments and suggestions, we would like to thank many seminar participants, David Malueg, Patrick Greenlee, Alex Raskovich, George Rozanski and, especially, Andrew Dick. We alone are responsible for the views expressed in this paper.

1. Introduction

Transactions through electronic payment networks (EPNs) in the U.S. exceeded $1.7 trillion in 2002 and are growing rapidly.[1] Several practices in this important industry have attracted controversy and antitrust scrutiny.[2] One such practice involves constraints on the ability of merchants to set different prices depending on the means of payment employed, such as credit or debit cards, cash or checks. We examine these constraints as instruments of vertical control, assess their welfare effects, and show that their presence may explain the phenomenon of rebates and reward programs in payment markets.

Uniform price constraints were at various times imposed by law or by EPN rules that prohibited merchants from imposing surcharges (or adverse non-price terms) for payments with an EPN card, even though merchants may face higher costs for card transactions due to fees charged by the EPN.[3] Even in the absence of formal prohibitions, merchants are often reluctant to set different retail prices for different means of payment.[4] We refer to all these limits as the No-Surcharge 'Rule' (NSR). Our analysis is relevant to assessing the desirability of laws or private regulations governing surcharging; for example, prohibitions on surcharges are banned in the United Kingdom and in Australia (Reserve Bank of Australia 2002). When such repeal of the

---

[1] Credit cards and offline debit cards accounted for $1.5 trillion ($755 billion for Visa and $444 billion for Mastercard — both known as bank card associations — and the rest through proprietary networks such as American Express and Discover) and $203 billion was via online debit cards. *Nilson Report*, March & April 2003, issues 784 & 785; ATM Debit News EFT Data Book 2003.

[2] For example, there is debate over whether the joint setting of certain network fees by EPN-member banks (as in bank card associations and regional networks) is anti-competitive (Salop 1990; Carleton and Frankel 1995, 1995a; Evans and Schmalensee 1995, 1999). Also, a common EPN requirement is that merchants must accept all of an EPN's cards (e.g. debit and credit cards) if they wish to accept any. This requirement was the target of a major lawsuit by Walmart and other retailers against Mastercard and Visa, alleging anti-competitive tying. Visa and Mastercard recently settled this suit, agreeing to relax the requirement and pay plaintiffs over $3 billion.

[3] Surcharges on credit card transactions were prohibited by federal statutes from 1968 to 1985 and remain prohibited by some states (e.g., Florida). For a detailed history of the U.S. legislative and regulatory treatment of surcharges, see Chakravorti and Shah (2001). In the U.S., Visa long had its own no-surcharge rule which it relaxed recently. Mastercard currently prohibits its merchants from "surcharging" customers for credit purchases, though it allows cash "discounts" (www.mastercard.com/consumer/cust_serv.html). In some European countries, card associations prohibit both discounts and surcharges. (Rochet and Tirole 2002.)

[4] Cash discounts are rare. According to one retailer survey, fewer than 1% of merchants offer cash discounts. Chain Store Age, *Fourth Annual Survey of Retail Credit Trends*, January 1994, section 2.

NSR is not an option — because merchants' reluctance to surcharge derives from other characteristics of the trading environment — our analysis helps understand, for example, the welfare effects of an EPN's granting of rebates to card users. Finally, the analysis is a necessary step towards evaluating card tying policies (see fn. 2 above), since such tying would have no force if merchant surcharges were unrestricted.

We consider a monopolist EPN that contracts with a representative merchant. The merchant faces downward-sloping demand from consumers, who pay with the EPN card or other means, 'cash'.[5] The EPN may set charges to the merchant and to card users. In this setting, one might expect an NSR to increase total consumer surplus and overall welfare, drawing on intuition from optimal taxation (or Ramsey pricing), where inefficiency is reduced by using a broader tax base to lower the tax rate. The analogy is that, for a given above-cost charge from the EPN to the merchant, an NSR leads the merchant to set an intermediate uniform price for all transactions instead of a higher price for card than for cash transactions; such price uniformity can reduce misallocation in the mix of transactions.

We show, however, that the optimal tax analogy is flawed. Here, the EPN is unregulated and allowing it to tax also non-card sales — indirectly via the NSR — leads the EPN to raise its charge to the merchant. As a result, the NSR can bring about higher merchant prices for cash *and card* transactions. This contrasts with standard comparisons of uniform pricing and third-degree price discrimination where, under regularity conditions also satisfied here, a requirement of uniform pricing causes at least some price(s) to fall (Nahata et al. 1990; Malueg 1992).

Our analysis also highlights a contrast between the NSR and a well-known instrument of vertical control, maximum resale price maintenance (RPM). Both practices enable a supplier to reduce the margin charged on its product by an imperfectly-competitive downstream firm. Maximum RPM, however, affects only the targeted product, lowering its price and benefitting consumers and overall welfare (Tirole 1988). In contrast, an NSR squeezes the merchant's margin indirectly, by requiring the same retail price to be charged for the other product (here, cash transactions), causing its price to rise.

---

[5] We use 'card' to denote any electronic payment instrument, and 'cash' to denote the alternative means of payment. Also, we sometimes will refer to the EPN as the card company. We abstract from the credit role of some electronic payments instruments and focus solely on its payment function. Chakravorti and Emmons (2001) present a model where some consumers use cards for both functions while others use them only as a payment instrument, and investigate the presence of cross-subsidies under an NSR from the former to the latter.

By constraining merchant pricing, an NSR alters the EPN's preferred allocation of fees. If merchant surcharges to consumers were unrestricted, only the EPN's total fee would matter, its division between card users and merchants would be irrelevant. With the NSR, however, the EPN concentrates its charges on merchants. Indeed, the EPN then prefers negative fees (rebates) to card users. Rebates are often viewed as stemming from the inability of an EPN to prevent its member banks from competing for card users and dissipating rents generated by high EPN fees to merchants. Our analysis reveals a different possible role: rebates allow an EPN to better exploit an NSR. Consistent with this interpretation, rebates have been offered not only by bank-card associations (such as Visa and Mastercard), but also by proprietary networks such as Discover, where a single entity controls the charges to both merchants and card users. In our model, the EPN grants rebates to boost card users' demand and raises its charge to merchants knowing they will absorb part of the increase, since the NSR requires that any increase must apply equally to cash users. Rebates thus can misallocate transactions towards cards, the opposite of what occurs absent an NSR, with ambiguous effects on overall welfare.

Despite the significant and growing importance of electronic means of payment (Evans and Schmalensee 1999), there are relatively few formal economic analyses of payment networks (for a survey of recent work, see Chakravorti, 2003). The standard reference on the operation of a payment system that involves multiple banks (bank card networks) is Baxter (1983). As Baxter illustrates, a typical payment transaction involves four parties. A card user exchanges with a merchant a promise to pay (a credit card receipt, for example) in return for goods. The merchant sells the promise to pay to a bank with which it has contracted, known as the merchant's 'acquiring bank'. The merchant receives the face value of the promise minus a fee called the 'merchant discount'. The acquiring bank then sells the receipt to the bank that issued the card ('issuing bank'), again at face value minus a fee known as the 'interchange fee'. The issuing bank collects from the cardholder the amount promised to pay. The issuing bank may also charge the card user a fee (or grant a rebate).

In a bank-card network, each member bank sets its own terms to card users or merchants. In *proprietary* networks such as American Express, Discover and Diner's Club, the same integrated entity deals with both merchants and card users. The interchange fee in this case is a fiction and there are only two true prices for EPN services, the merchant discount fee and the card user fee. Our model considers a profit-maximizing agent, the EPN, setting the charge to a

merchant and, for most of the paper, also to card users. This model is most obviously interpreted as one of a proprietary card network. It also characterizes the behavior of a bank card network under two hypothetical conditions: (a) acquiring banks are identical and competitive; and (b) issuing banks are identical and collude in pricing to card users. Condition (a) implies that variations in the interchange fee are fully passed through to the merchant discount, and the merchant discount is effectively set by the EPN's issuing banks through their choice of interchange fee. Condition (b) implies that card user charges are chosen to maximize overall profits of issuing banks.[6] In Section 6 we analyze the other polar case of competition among issuer banks — card user fees are then set in a Bertrand fashion.

The complexity of this market implies that EPN pricing practices will influence multiple decisions. For example, merchants must decide whether to accept a card, and customers must choose whether to use a card and, also, their level of purchases. Schmalensee (2002) examines how the joint setting of interchange fees affects the marginal decisions of merchants to accept cards as well as consumers' decision to carry cards. Rochet and Tirole (2002) analyze the impact of interchange fees and the NSR on the decision of a consumer to use a card versus an alternative means of payment. In their model, consumer have unit demands for a good but differ in their private values of paying with cards versus cash. The net cost of using cards affects the number of consumers who choose cards rather than cash. However, as Rochet and Tirole observe, the total quantity of purchases is unaffected by the NSR given the assumptions of unit demands and that all consumers are served in the Hotelling competition among merchants.

Our focus is on the impact of an NSR on the *quantities* of purchases. We assume that the means of payment for a given consumer is exogenous; a fraction of consumers use cards and the rest use cash. In our model, however, consumers have continuous demand functions for goods. Thus, while EPN pricing and the NSR will not affect the number of card users in our model, it can and will affect the total quantities of card and cash transactions.[7]

---

[6] Rochet and Tirole (2002) and Schmalensee (2002), discussed further below, assume imperfect competition at the merchant level and in card issuance, but perfect competition in merchant acquisition (citing empirical evidence that acquisition in the U.S. is significantly more competitive than issuance).

[7] Wright (2000) uses the Rochet and Tirole framework to analyse the role of the NSR in preventing ex post opportunistic pricing against cardholders by a monopolist merchant. Other differences between our approach and Rochet and Tirole's are less consequential. For example, they assume (as do we) that EPN members set the interchange fee (and ultimately the merchant discount) jointly, but may compete in their fees to cardholders. This is

The paper is organized as follows. Section 2 presents our model without the NSR, where a monopolist EPN can commit to setting linear charges to card users and to a monopolist merchant, that in turn can set different prices to cash and card users. The relative size of these groups is a key parameter of the model; another is the merchant's gross benefit from card rather than cash transactions. Section 3 shows that with the NSR, the EPN strictly prefers to maximize its charge to the merchant and minimize its charge to card users. Section 4 considers the case where rebates to card users are not feasible. With a small enough cash market, the NSR harms not only the merchant and cash users, but even *card* users. With a larger cash market, card users gain but, for linear demand, aggregate consumer surplus of cash plus card users is always lower than without the NSR. Total surplus, however, is higher if the cash market is sufficiently large.

Section 5 allows for rebates. Given the complicated constraint sets, we specialize the analysis to linear demand. Rebates enable the EPN to fully extract the merchant's surplus in some cases where it could not without rebates, but allows this extraction to occur in a way that yields higher total (card plus cash) transactions. Under the NSR, the use of rebates benefits card users, aggregate consumer surplus, and total surplus, though cash users lose. Compared to no NSR, the NSR with rebates increases total surplus if the cash market is sufficiently large and increases overall consumer surplus if the cash market is sufficiently *small*. A larger merchant benefit from card transactions improves the effects of the NSR on both total surplus and overall consumer surplus — the reverse of what occurs when rebates are not feasible.

Section 6 modifies the industry structure by assuming that the EPN cannot control the prices that member banks charge card users. An NSR again will induce rebates, benefitting card users but harming cash users. If competition among the member banks is strong (Bertrand), then — for linear demand — an NSR increases overall consumer surplus regardless of the relative sizes of the two consumer groups. However, as long as the merchant's benefit from card use is sufficiently low, the NSR reduces overall welfare.

---

modeled by treating the equilibrium cardholder fee as a reduced form function that decreases in the interchange fee. (Raising the interchange fee raises a cardholder bank's margin, thereby inducing the bank to cut its cardholder fee so as to expand card usage.) The two cases we consider — an EPN monopolist that sets cardholder fees directly, or Bertrand competition among EPN members — are special cases of their specification. Regarding merchant behavior, they assume symmetric duopolists in Hotelling competition while we assume a monopolist, but in both cases the number of merchants is fixed and each faces downward-sloping demand. Like Rochet and Tirole, we address whether the merchant will agree to accept a card.

2. The Model and Pricing With Surcharging

Consumers: We consider two types of consumers. Type *e* consumers ('card users') hold cards from the EPN. They buy units of a good only by using cards; their mass is *1*. Type *c* consumers buy units of a good using only an outside means of payment, call it cash. We assume that they do not have cards; their mass is α. The presence of a sizable portion of the population who *must* pay by cash because they are ineligible to acquire cards is plausible.[8] The assumption that cardholders cannot use cash is, of course, more problematic. At the end of this section, we suggest a model with endogenous choice of payments that mimics the effects of this assumption. The advantage of this simplifying assumption is to allow us to conduct simple comparative statics by using α as an index of the relative importance of cash and cards in the market.

Consumers are otherwise identical and have quasilinear preferences from purchases of goods given by

$$U(p_c, q_c)=V(q_c)-p_c q_c,$$
$$U(p_e, q_e)=V(q_e)-p_e q_e,\ V'(\bullet)>0,\ V''(\bullet)<0.$$

Throughout, q_j is the *per capita* number of transactions of a consumer of type $j = c, e,$ and p_j is the net price per unit of transaction paid by such a consumer. The net price paid by cash users equals the price charged by the merchant but the two prices may differ for card users: $p_e = p_e^M + t,$ where p_e^M denotes the price charged by the merchant to a card-using consumer and t is the per unit charge (or rebate if $t < 0$) imposed by the EPN on card users. For each type of consumer, the (downward sloping) inverse demand function is given by $V'(q_j) = p_j$.

Merchants: We assume that merchants are local monopolists who treat the above inverse demand curve as the demand for their product from each type of consumer. The marginal cost of providing a good to a cash consumer is assumed constant and is normalized to zero. The merchant may also gain a benefit, $b \geq 0,$ from being paid by card instead of cash, reflecting potential savings on cash-handling costs. The size of b affects the degree of distortion due to double marginalization under surcharging. The merchant is charged a per-unit fee i by the EPN.

---

[8] About 24% of U.S. families do not hold cards of any kind (Federal Reserve Board, 2001, p. 25). Presumably a large fraction of these families cannot get cards. Evans and Schmalensee (1999) characterize non-card holders as being "on the economic fringes of society" (p. 87), with a median household income 50% below the overall average, and more than 40% of them with incomes below the government's estimated poverty line.

The merchant's profit is $\alpha p_c q_c$ from cash users and $p_e^M q_e - (i-b)q_e$ from card users, where quantities are given by $p_c = V'(q_c)$ and $p_e^M = V'(q_e) - t$. For given values of i and t, the merchant's problem can therefore be formulated as choosing a level of x to solve

$$max_x (V'(x)-(i+t-b))x.$$

Observe that $i=t=b=0$ yields the merchant's problem vis-a-vis the cash market. Written this way, the term $i+t$ can be interpreted as a tax imposed on the card market by the EPN. For given (i,t), we denote the value of this optimization problem by $\Pi^M(i,t;b)$.

The merchant's alternative to accepting cards is to serve the cash market alone, yielding a profit-maximizing per-capita level of transactions x_0 and total profit $\alpha x_0 V'(x_0)$. The merchant must be assured at least this amount in any equilibrium, that is, for any (i,t),

(IR) $\Pi^M(i,t;b) \geq \alpha x_0 V'(x_0)$.

We refer to this as the 'individual rationality' or IR constraint.

Electronic Payment Network: As noted, we suppress the distinction between the interchange fee and merchant discount, and simply view the EPN as setting the charge to merchants, i, monopolistically. With the exception of Section 6, we also assume that the EPN acts monopolistically in the setting of any card user fees. The timing of price setting is in a Stackelberg manner: that is, the EPN sets t and i and commits to this profile of prices and, given t and i, the merchant sets her monopoly price. The EPN's marginal cost of servicing a card transaction is assumed to be zero.

We assume that two-part tariffs are not available either to the EPN or to the merchant. What is important for our analysis is that the sequential monopoly environment between the card company and the merchant lead to some inefficient pricing at both the merchant and EPN levels.[9] For simplicity, we assume that only linear pricing is feasible for each agent.

The first-order conditions from the merchant's problem yields a derived inverse demand

---

[9] There are a variety of reasons why fully efficient two-part tariffs (or other nonlinear pricing) may not be achievable for the EPN to eliminate such double marginalization. A typical EPN has relationships with a vast number of merchants, and contracting costs could make merchant-specific, two-part tariffs prohibitively expensive. Furthermore, merchants aggregated together in a single market place, such as a mall, may be able to avoid most of the impact of a fixed fee by channeling all card purchases to a single merchant. Additionally, in the context of asymmetric information, for example with heterogeneous merchants, the optimal two-part tariff generally yields some surplus to the high demand merchant and pricing at levels above marginal cost.

curve for card transactions defined, implicitly, by

$$i+t=V'(x)+xV''(x)+b \tag{1}$$

Therefore, the EPN maximizes $(i+t)\,x$ or

$$\Pi^{e}(b)= \max_{x} (V'(x)+xV''(x)+b)\,x \tag{2}$$

Since x is a function of $i+t$ but not i or t separately, the card company varies x by varying the sum of charges, $i+t$. This leads immediately to the following well-known result.[10]

Proposition 1: *Suppose merchant surcharges are allowed. If (i,t) maximizes the profits of the EPN, then so too does any pair (i′,t′) where* $t'+i'=t+i$.

That is, it is irrelevant whether the EPN charges its fee to the consumers, to the merchant, or to a combination of the two. Since the sum, $i+t$, can be viewed as a transactions tax, Proposition 1 echoes the familiar result that the effects of a tax are invariant to whether the obligation to pay the tax is placed on buyers or on sellers. However, the next section shows that, in the presence of an NSR, EPN profits will vary for a given $i+t$ depending on the relative values of i and t.

The remainder of the paper imposes some further restrictions on the environment:

A1) *i) The merchant's cash market revenue function, p f(p) where f(•) is the inverse of V′(•), is concave in price; ii)* $x\,V'''(x)+V''(x) \leq 0$ *which implies the revenue function is also concave in quantity*[11]*;*

A2) *The EPN's revenue function* $x(x\,V''(x)+V'(x))$ *is concave in quantity;*

A3) *For* $x_0 = \operatorname{argmax}_x\, x\,V'(x)$, $b < -x_0^{\,2}\,V'''(x_0) - 2\,x_0\,V''(x_0)$.

The first two assumptions are concavity assumptions which are sufficient to ensure a unique solution to the various optimization problems that arise in the market where surcharging

---

[10] This result was noted in Carleton and Frankel (1995). A generalization of the result and an explanation of the intuition underlying it can be found in Gans and King (2003).

[11] Throughout, the concavity conditions apply only over the region where quantities and prices are positive.

is feasible.[12] Assumption **A3)** implies that the merchant's benefit from card use (b) is not so great that — in spite of the double-marginalization on cards — card transactions would exceed cash transactions. (See Lemma 2.)

Note that all three assumptions are satisfied under linear demand. Given constant, zero marginal costs, we can assume a unit intercept and slope without further loss of generality, so throughout we use 'linear demand' to mean the case

$$V'(x) = 1-x,\ x \in [0,1].$$

Here, **A3)** is satisfied if $b < 1$, that is, under the mild condition that the merchant's gross benefit from cards is no higher than the choke price of cash users.

The subsequent analysis uses the following lemma frequently. Define $x(k)$ as $x(k)=argmax_x (V'(x)-k)\ x$, the profit-maximizing quantity for a monopolist that faces inverse demand $V'(x)$ and marginal cost k. Assumption **A1)** ensures that the profit-maximizing quantity is unique, though this is not required for the Lemma.

Lemma 1: *$k'>k$ implies $x(k')<x(k)$.*

The proof, which is in the Appendix along with all other proofs, uses a standard revealed preference argument (see, e.g., Tirole 1988, pp. 66-67). Note that for a given i and t, the per capita quantity of card transactions is $x(i+t-b)$ while the cash quantity is $x(0)$. Lemma 1 implies that when the sum of the fees to the merchant and card user exceeds the merchant's benefit from card use, $i+t > b$, then per capita card transactions are lower than cash transactions.

When is it optimal for the EPN, incorporating the merchant's behavior, to choose a total charge that exceeds the merchant's benefit from card use, $i+t>b$, and thereby bring about a higher net price for card transactions than for cash? Lemma 2 exploits equations (1) and (2) to show that this will occur if the merchant benefit from card use is not too high.

Lemma 2: *Assume **A1)-A3)**. In the environment with merchant surcharges allowed, per capita card use is less than per capita cash use.*

---

[12] Assumption **A1)ii)** is, of course, slightly stronger than concavity. It implies that if a monopolist with a revenue function $x\ V'(x)$ incurs an increase in his (constant) marginal cost, this increase is not fully passed on in price to consumers.

Before proceeding, we discuss briefly the assumption that a consumer's means of payment is exogenously fixed. While it is plausible that some fraction of customers simply cannot qualify for cards, two issues remain: 1) when surcharging is possible, why do card users not switch to cash to avoid the higher card price?; 2) if the merchant were to refuse cards, why does she lose all cardholding customers (as implied by the IR constraint formulated above) instead of switching at least some of them to cash? Both issues can be resolved by nesting our model within an extended model where the means of payment can be chosen endogenously but cardholders continue to behave according to our assumptions.[13] For simplicity, we do not make this extension explicit because the real question is this: can the EPN still tax cash users once the choice of payment mode is endogenized? Rochet and Tirole (2002) illustrate that the answer is yes, when cash and cards are imperfectly substitutable.[14] Besides the advantage of simplicity, our model allows us to utilize the relative sizes of the two groups, α, to conduct comparative statics.

---

[13] Suppose that card users obtain a fixed benefit, B, from the use of cards rather than cash. (Their utility is given by $U(p_e, q_e)=V(q_e)-p_e q_e+B$). As an alternative to purchases from the merchant, cardholders can buy from an outside, card-serving market which is competitive but higher cost and charges a fixed price, p^* (inclusive of card fees). Cardholders purchase from the local merchant offering card price p_e if

(C1) $p^* > p_e$.

Define q^* such that $V'(q^*)=p^*$. Cardholders purchase from the competitive fringe rather than paying with cash to the local merchant charging p_c if

(C2) $V(q^*)-p^*q^*+B \geq V(q_c)-p_c q_c$.

As long as B is sufficiently high and p^* is sufficiently high but not too high, the two conditions can be met. The local merchant would lose cardholders to the fringe if it dropped cards (**C2**) and, by accepting cards, will retain these customers even when setting its monopoly card price under surcharging (**C1** and **C2** jointly).

[14] Using the notation of Rochet and Tirole (RT), under no NSR the Hotelling duopolist merchants charge identical equilibrium prices to cash users, $p_{cash}^*= d + t$, where d is the merchant's marginal cost of providing the good excluding payments to the EPN and t is the transport cost parameter in the Hotelling model. Let c_A and c_I denote the resource marginal cost to an issuing and an acquiring bank, respectively, and a denote the interchange fee paid to an issuing bank. Under the NSR the equilibrium price charged by both merchants is given in RT's equation (5), $p^*=[d+D(f^*(c_I-a))m^n(a)]+t$, where $D(.)$ is the demand for cards as a function of the equilibrium cardholder fee, f^*, and $m^n(a)$ is the merchant's net marginal cost of card transactions. Thus, $m^n(a) =c_A+a-b_s$ where c_A+a is passed on to the merchant by the competitive acquiring banks via the merchant discount, and b_s is the merchant's gross benefit of processing card versus cash transactions. Observe that $p^* > p_{cash}^*$ (cash users are taxed) if $m^n(a)>0$, which holds under the mild assumption that the potential card user benefits are non-negative.

3. Under a No Surcharge Rule the EPN Prefers Lower Card User Charges

Suppose the EPN requires any merchant that accepts its card to charge no more to card users than to cash users, $p_e^M \leq p_c$.[15] In this section we demonstrate that such a pricing constraint binds on the merchant whenever the EPN's fee to card users is low enough. We also show that in contrast to the neutrality result when merchant surcharging is allowed, with a binding NSR the EPN prefers to offer a high merchant fee and a low card user fee.

Proposition 1 showed that, with surcharging allowed, EPN profits are constant for any given $i+t$. Nevertheless, the prices that a merchant charges to different consumers will vary depending on how the EPN divides its aggregate 'tax' between the merchant and card users. Lemma 2 showed that the EPN's optimal aggregate fee with no NSR will satisfy $i+t > b$, so the merchant faces a higher net marginal cost of serving card users than cash users; thus, if $t = 0$ a card user's inverse demand is equal to that of a cash user, hence the merchant's higher marginal cost dictates setting a higher price to card users ($p_e^M > p_c$). Lemma 3i) below shows that the same effect occurs if, instead, the EPN levies on cardholders a sufficiently small positive charge.[16] Lemma 3ii) shows that, holding constant the sum of the EPN's fees, $i+t$, when t is low enough that the NSR would bind, imposing the NSR lowers cash purchases but raises card purchases, thereby benefitting the EPN. Finally, Lemma 3iii) shows that, under an NSR, the EPN has the incentive to raise the charge to the merchant and lower the charge to card users.

Lemma 3: *Assume **A1)-A3).** Fix $k>b$ and define $t^*(k) \equiv V'(x(k-b))-V'(x_0) >0$. For any (i,t),with $i+t=k$, $t < t^*(k)$:*

i) When merchant surcharges are allowed, $p_e^M > p_c$ implying that with this profile of fees the imposition of an NSR constrains merchant pricing;

---

[15] Of course, a merchant may refuse and forgo card transactions. To understand the direction of EPN incentives under the NSR, in this section we examine the structure of EPN pricing assuming the merchant's IR constraint does not bind. In later sections we address this constraint explicitly.

[16] An implication of this observation is that, with a low cardholder fee (which we show is desired by the EPN), we can formulate the no-surcharge rule mathematically as the inequality constraint, $p_e^M \leq p_c$ even if, formally, the constraint is a 'No Discrimination Rule', that is, a uniform pricing rule rather than a no-surcharge on card use rule (which would be better captured by the constraint, $p_e^M = p_c$). Although credit card companies, for example, have historically imposed such rules on their merchant clients, an inequality constraint may obscure other reasons for merchant pricing constraints. Some merchants argue that even without a formal no-surcharge rule, social conventions make it very difficult for them to charge different prices for users of different means of payments. Lemma 3i) shows when the effects of the two constraints are the same.

ii) If an NSR is accepted, holding (i,t) fixed, then card purchases rise and cash purchases fall; iii) Fix any (i,t) and (i′,t′), with $i+t=i'+t'=k$, $t' < t < t^*(k)$. As long as the merchant accepts the NSR with both profiles of charges, per capita card purchases and EPN profits are higher with (i′,t′) than with (i,t).

The intuition for Lemma 3i) is straightforward. Suppose the EPN sets *t* low enough and *i* high enough that the merchant charges $p_e^M > p_c$. The NSR then binds and induces the merchant to choose a uniform price between these two levels; starting from a uniform price equal to p_e^M, a small move towards p_c imposes a zero first-order loss in the card market while moving closer to the optimal cash price, and similarly starting from p_c^M and moving towards p_e^M.[17]

Lemma 3ii) suffices to establish that the EPN's profit rises when a binding NSR is accepted, since the EPN's profit increases at the pre-NSR charges, *(i,t)*. By revealed preference, any departure from these charges post NSR would further increase profit. In which direction would the EPN alter its charges? Low card user fees (*t*) ensured that the NSR was binding on the merchant. Lemma 3iii) shows that the EPN benefits by choosing a small reduction in *t* accompanied by an equivalent increase in *i*, since this further increases card transactions.

The intuition behind Lemma 3iii) is as follows. A cut in *t* and an offsetting increase in *i* would leave the EPN's margin unchanged, and hence profit unchanged, only if card transactions remained unchanged. This in turn would only happen if the merchant raised her price to card users by the full increase in *i*, since card users' inverse demand shifts up by an amount equal to the fall in *t* (equivalently, to the increase in *i*). But since the NSR forces the merchant to charge the same price to cash users as to card users, and since the marginal cost of serving cash users has not changed, the merchant prefers to raise its uniform price by less than the full increase in *i* and accept a lower margin on card sales.[18] Interestingly, a stronger sufficient condition is required to

---

[17] This result echoes the finding prohibiting third-degree price discrimination leads a monopolist to charge an *intermediate* uniform price. Sufficient conditions are that marginal cost be non-decreasing and that demands in the various markets be independent, each yielding a quasi-concave profit function (Nahata et al. 1990, Malueg 1992).

[18] This argument implies that, under the NSR, the total price to card users falls by more with a unit reduction in *t* than in *i*. Let *p* denote the merchant's price to card users, hence their total price is $p+t$. Under surcharging, Proposition 1 implies that a unit reduction in *t* or in *i* yields the same cut in $p+t$: $\partial(p+t)/\partial t=\partial(p+t)/\partial i=\partial p/\partial i$, so $1+\partial p/\partial t=\partial p/\partial i$. The NSR, however, dampens the merchant's price response to a cut in *t* or in *i* ($|\partial p/\partial i|$ and $|\partial p/\partial t|$ fall), because the same price change must be made also in the cash market. Thus, cutting *i* yields a smaller reduction in the price to card users under the NSR than under surcharging ($\partial p/\partial i$ is smaller), while cutting *t* yields a larger reduction than under surcharging: card users receive this cut directly, and the merchant responds by increasing *p*

show that cash transactions decline as t falls.[19] However, we show later that the EPN's profit-maximizing prices (i, t) under the NSR indeed cause cash transactions to decline.

Lemma 3 shows that with the NSR, it becomes relevant how $i+t$ is distributed: the EPN prefers lower t (provided the merchant still accepts). This provides an alternative explanation for why rebates (negative t) are offered to card users. Such rebates are often taken as evidence of the inability of a bank card association to control competition for card users by its member banks. Lemma 3iii) offers an alternative interpretaion: rebates may be a pricing tactic designed to better exploit the power of the NSR.[20]

Given the incentives for an EPN to raise i and reduce t, what determine the floor on t? One limit may be institutional. For historical, practical or other reasons, rebates may not be an option for the EPN.[21] Section 4 investigates the effects of the NSR when rebates are not possible. In this case, the binding constraint may be the non-negativity of t, the merchant's option to reject the NSR and serve only cash users, or both.[22] In fact, as Proposition 2 illustrates, the non-negativity constraint always binds. Section 5 allows for rebates. In this case, we show that the EPN may be constrained either by the IR constraint or by the constraint that the merchant continue to be willing to serve the *cash* market (a sort of incentive compatibility constraint).

(given assumption **A1)ii**) by less under the NSR.

[19] A proof along the same lines as that for Lemma 3iii) shows that $V'(\bullet)$ convex is a sufficient condition.

[20] Gerstner and Hess (1991) obtain a similar effect in a somewhat different context. They consider a monopolist manufacturer selling to a monopolist retailer that faces two customer groups, low demanders and high demanders, where high demanders incur a higher transaction cost of using a rebate/coupon. In our model, the NSR plays roughly the same role as their differential transaction costs in motivating rebates.

[21] The phenomenon of card user rebates is relatively recent. While credit cards date to the late 1960s/early 1970s, money-back rebates were first offered, by Discover, in 1986. Rebate cards only became common, however, in the early 1990s with the introduction of the GM Mastercard and other cards that offer reward points associated with co-branding partner companies (such as frequent-flier miles). See generally, Evans and Schmalensee (1999). Today, roughly half of all credit volume is associated with rebates of various sorts. Faulkner and Gray (2000).

[22] In (i,t) space, under the NSR the merchant's level sets have slope strictly less than -1. Therefore, for any given k, the line $t=k-i$ eventually crosses the line given by $\Pi^{NSR}(i,t;b)=\alpha\, x_0 V'(x_0)$. Thus, if the EPN holds $i+t$ fixed and lowers t, it eventually runs against the merchant IR constraint.

4. Equilibrium Under No Rebates

Let i_0 be the EPN's optimal charge given $t = 0$ and (for the moment) ignoring the merchant's IR constraint. Whether or not the IR binds depends on the relative size of the cash market, α. If, at $(0,i_0)$, the IR does not bind, then, by Lemma 3, these prices are optimal for the EPN. If the IR is violated at these prices, in Proposition 2 we provide sufficient conditions under which setting $t = 0$ is still optimal for the EPN.

The analysis in Propositions 2 and 3 of the case where the merchant's IR binds requires an additional assumption that enables us to focus solely on the first order conditions of the EPN's optimization problem under an NSR.[23] Define the merchant's profit function and the merchant's profit maximizing choice of card transactions under an NSR as

$$\Pi^{NSR}(i\ t;b) \equiv max_x\ \alpha f(V'(x)-t)\ (V'(x)-t)+ x\ (V'(x)-i-t+b),$$

$$q_e(i,t;b) \equiv argmax_x\ \alpha f(V'(x)-t)\ (V'(x)-t)+ x\ (V'(x)-i-t+b).$$

The EPN's profit maximization problem assuming an NSR is imposed can be expressed as

$$P^{EPN}: \qquad max_{i,t}\ (i + t)\ q_e(i,t;b)$$

$$s.t. \qquad \Pi^{NSR}(i,t;b) \geq \alpha\ x_0\ V'(x_0) \qquad (IR)$$

$$t \geq 0 \qquad (No\ Rebates).$$

We have not found transparent general conditions that would ensure that this optimization problem is a concave program for all α, however, it is clear that, for low α, the IR constraint does not bind. In this case, Lemma 3 and Assumptions **A1-A3)** imply that the first order necessary conditions for an optimum are also sufficient. Similarly, for values of α slightly above the level of α that the IR binds, first order necessary conditions are sufficient. Assumption **A4)** is a technical assumption that ensures the first order necessary conditions are also sufficient for an optimum no matter how large is the cash market.

A4) *For all α, if (i^*,t^*) satisfy the Kuhn-Tucker first order conditions for P^{EPN}, then (i^*,t^*) solve P^{EPN}.*

The case of linear demand satisifes **A4)** for all values of α.

---

[23] Proposition 3i) does not rely on **A4).**

Proposition 2 (Prices): *Assume **A1)-A4)** and suppose card rebates are not feasible ($t \geq 0$). Under the NSR:*

i) The EPN's optimal fee implies $t = 0$ (no card fees), hence per capita card and cash transactions are equal.

ii) There exists α^* such that the EPN choice of i is determined by the merchant's IR constraint if and only if $\alpha > \alpha^*$.

iii) If $\alpha > \alpha^*$, $i+t-b$ $(=i-b)$ is independent of b and falls as α rises.

iv) Under linear demand, i is higher than the total charge under surcharging for all α; i increases in α for $\alpha < \alpha^*$, otherwise it decreases in α.

For Proposition 3ii)c), define the change in total surplus when rebates are not feasible (ΔTS^{NR}) to be total surplus under the NSR without rebates minus total surplus under no NSR.

Proposition 3 (Quantities and Welfare): *Assume **A1)-A4)** and suppose an NSR is imposed but card user rebates are not feasible ($t \geq 0$). Compared to the equilibrium with no NSR,*

i) For $\alpha \leq \alpha^*$ (IR does not bind):

a) Cash users' transactions and consumer surplus are lower;

b) Card users' transactions and consumer surplus are unchanged if $b = 0$ and lower if $b > 0$;

ii) For $\alpha > \alpha^*$ (IR binds):

a) Cash users' transactions and consumer surplus are lower;

b) Card users' transactions and consumer surplus are higher if α is sufficiently larger than α^*;

c) Under linear demand, aggregate quantity $(q_e+\alpha q_c)$ and aggregate consumer surplus are lower for all and all b. ΔTS^{NR} rises in α and falls in b. For $b=0$, $\Delta TS^{NR}=0$ at a value of α above α^*.

Proposition 2i) shows that the EPN's desire for lower card user fees and higher merchant fees illustrated in Lemma 3 is stronger than the merchants' preference for the reverse pattern. Even if the merchant's IR constraint binds on the EPN before the EPN achieves its optimal fee pair, the EPN will choose to move down the merchant's IR locus to set card user fees to zero: the EPN chooses its unconstrained optimum $(i_0,0)$ if the merchant's IR is not binding at this point,

otherwise it accepts a merchant fee lower than i_0 but maintains $t=0$. An implication is that if the non-negativity constraint is relaxed (rebates are allowed, as in Section 5) then the EPN under the NSR will set t negative.

Proposition 2ii) shows that the merchant IR constraint binds if and only if the cash market is not too small. (With linear demand and $b=0$, the IR binds if $\alpha > \alpha^* = 1/3$.) Intuitively, the merchant's profit from serving only cash customers is proportional to the size of the cash market, therefore, as the latter increases, the EPN eventually must depart from its unconstrained optimal charges to maintain merchant participation.

IR Not Binding. When $\alpha < \alpha^*$, the welfare consequences of the NSR are stark. The NSR reduces even card transactions (leaving them unchanged only if $b=0$), thus harming card users. (Proposition 3i)). Since (per capita) cash transactions exceed card transactions under surcharging but are equal to them with the NSR, the NSR also reduces cash transactions. With all quantities falling, total surplus must fall. The merchant's profit also falls since the NSR both leads to a higher total EPN charge and constrains the merchant's pricing to consumers. The NSR in this case therefore benefits only the EPN at the expense of all other parties.

One might have expected the NSR to raise card transactions by inducing the merchant to choose a uniform price that lies between its card and cash prices under surcharging. Instead, the NSR leads the EPN to adopt a merchant fee i_0 so much higher than its total fee $i+t$ under no NSR that card transactions remain unchanged with the NSR if $b=0$ and fall if $b>0$. [24]

IR Binding. When the cash market is large enough that the merchant's IR constraint binds on the EPN's pricing, the NSR still reduces the merchant's profit — since the merchant

---

[24] The quantity effects can be understood as follows. Under surcharging, the derived inverse demand function facing the EPN is $i(q) = b + (V''q + V')$, where $V''q + V'$ is the merchant's decreasing marginal revenue function. With the NSR, the EPN faces $i^N(q) = b + (1+\alpha)(V''q + V')$. Recall that x_0 is the merchant's monopoly output for zero marginal cost (the cash-market output with surcharging), hence: $V''(x_0)x_0 + V'(x_0) = 0$. Thus, $i^N(q)$ cuts $i(q)$ from above at $q=x_0$, $i=b$: $i^N(q) = i(q) = b$ at $q = x_0$, while $i^N(q) > i(q)$ at $q < x_0$ and $i^N(q) < i(q)$ at $q > x_0$. (Intuitively, $i=b$ makes the merchant's net marginal cost of card transactions zero—as for cash—so the merchant would choose equal card and cash quantities x_0 under surcharging hence the NSR would have no effect. Card quantities $q < x_0$ correspond to the merchant facing higher marginal cost for card than for cash sales, hence the EPN can attain such quantities at a higher i under the NSR because the merchant's uniform price is then pulled down by the lower marginal cost on cash sales; conversely, $q > x_0$ requires cutting i below b by more under the NSR than under surcharging.) The equilibrium card quantity is where the EPN's marginal revenue obtained from the relevant inverse demand function, $i^N(q)$ or $i(q)$, equals its marginal cost of zero: $MR = i(q) + i'(q)q = 0$ under surcharging, and $MR^N = i^N(q) + i^N{}'(q)q = 0$ under the NSR. Observe that $MR^N = (1+\alpha)MR - \alpha b$. Thus, for $b = 0$, the card quantity is the same with surcharging or the NSR. For $b > 0$, $MR = 0$ implies $MR^N < 0$, so the EPN's optimal card quantity is lower under the NSR. (Note that this will be true also if the EPN's marginal cost were positive but not too large.)

now loses all its surplus from dealing with the EPN — and cash transactions. However, if the cash market is sufficiently large, card transactions are higher with the NSR (Proposition 3ii)b)). To see this, observe that with the NSR and no rebates ($t=0$) the merchant's profit can be expressed as $-(1+\alpha)Q^2V''(Q)$, where Q is the equal per-capital level of cash and card transactions. The IR constraint is therefore $-(1+\alpha)Q^2V''(Q) = \alpha x_0V'(x_0)$ or

$$-Q^2V''(Q) = (\alpha/(1+\alpha))x_0V'(x_0). \tag{3}$$

As the size of the cash market, α, increases, to satisfy (3) the EPN must induce an increase in Q (since merchant profit is increasing in Q by **A1)ii),** concavity of the merchant's revenue function in quantity), which requires cutting the merchant fee i. As $\alpha \to \infty$, $(\alpha/(1+\alpha))x_0V'(x_0) \to x_0V'(x_0)$, so Q must approach x_0, the merchant's cash market quantity under surcharging. The NSR therefore lowers cash transactions (since $Q < x_0$ except in the limit), but for sufficiently high α it raises card transactions (since these are less than x_0 under surcharging, by Lemma 2).

The effect on total quantity depends on the precise form of the merchant IR constraint which, in turn, depends on $V(\bullet)$. Proposition 3ii)c) shows that for linear demand, total quantity is lower under the NSR. Given equal per-capita linear demands by card and cash users, imposing the NSR would leave total quantity unchanged only if the EPN's total charge remained unchanged, but in fact the EPN raises its total charge (Proposition 2iv) in order to exploit the decreased elasiticity of demand that it faces from the merchant, so total quantity falls. Overall consumer surplus of card and cash users must fall, because it would have fallen even if total quantity had remained constant. To see this, observe that per capita quantities are unequal under surcharging but equal under the NSR, and note the ensuing principle which, for future reference, we denote the 'trapezoids argument': Given identical, linear per-capita demands of cash and card users, any decrease in the gap between per capita quantities while holding total quantity fixed will reduce overall consumer surplus — the loss to the higher-quantity group from the rise in its price exceeds the gain to the other group from the fall in its price.[25] Since the NSR with no rebates reduces both the spread in per capita quantities and total quantity, overall consumer surplus must decline.

Total surplus, however, can be higher with the NSR if the cash market is large enough.

---

[25] Consider any two pairs of prices, $(p_e,p_c),(P_e,P_c)$ such that total quantity transacted, $q_e + \alpha q_c$, is the same. Then linear demand implies $p_e + \alpha p^c = P_e + \alpha P_c$. Consumer surplus is convex in price, thus if $|p_e - p_c| > |P_e - P_c|$ ($=0$ in the no rebate case with an NSR), aggregate consumer surplus is higher under the price profile (p_e,p_c).

The efficiency gain comes because the lower total quantity of transactions is allocated more efficiently between cash and card users. To see this, consider $b = 0$, in which case the welfare maximizing allocation requires equal per capita card and cash quantities. The NSR achieves this, while surcharging does not. If the cash market is sufficiently larger than the level where the merchant's IR binds on the EPN (for $b = 0$, if $\alpha > 1.53 > \alpha^* = 1/3$), then the benefit from improved allocation outweighs the harm from the reduction in total quantity so the NSR raises total surplus. Intuitively, a large cash market allows the NSR to curb the double marginalization that curtails card transactions under surcharging while introducing only a small distortion in the *per capita* quantity of cash transactions.

Finally, consider the role of b, the merchant's gross benefit from card rather than cash transactions. With linear demand and surcharging, the distortion from double marginalization on card transactions increases with b (the gap between the efficient and actual card quantities is $3(1+b)/4$). Since in our model double marginalization creates the motive for the NSR, one might expect the NSR to have a more favorable effect on total surplus the larger is b. In fact, the NSR's advantage over surcharging in raising card transactions is *smaller* the larger is b, hence for total surplus, the best case for the NSR occurs when $b = 0$. Section 5 shows that this conclusion is reversed when rebates to card holders are feasible; there, ΔTS is *increasing* in b.

5. Equilibrium When Rebates Are Feasible

Proposition 2i) shows that when the card user fee must be non-negative, the EPN cuts this fee to 0. Thus, this is no longer the equilibrium when rebates are feasible ($t < 0$).

One obvious constraint on the EPN's equilibrium charges remains the merchant's IR, its option to reject the EPN and forgo card users as discussed earlier. In addition, a less evident constraint emerges when rebates are feasible: the merchant's willingness to continue serving *cash* customers. With large enough card user rebates and a sufficiently small cash market, the monopoly price appropriate for card users alone will exceed the choke price of cash users, and the merchant under the NSR will choose this price instead of cutting price enough to serve also cash users.[26] Such an outcome, however, clearly is not optimal for the EPN: since the merchant's price to card users is then unaffected by cash users, the NSR loses its value. This issue does not

[26] A monopolist that faces two markets but is prohibited from 3rd-degree price discrimination will choose to serve only one market if the dispersion in the demands is sufficiently large (Tirole 1988, p.139).

arise with $t \geq 0$ — (per capita) inverse demand of cash users is then lower than that of card users, so any price that yields cash sales also yields card sales — but must be tackled under rebates.

General results are not available for the case of rebates because of the complicated nature of the constraint sets, so Propositions 4 through 6 below restrict attention to linear demand. Propositions 4 and 5 compare the equilibrium under the NSR with rebates to that under no NSR. Proposition 6 summarizes the incremental effect of rebates by comparing the equilibria under the NSR with and without rebates.

Proposition 4 (Prices): *Assume linear demand. Under an NSR with rebates feasible:*
i) For all α *, the EPN's optimal choice involves granting rebates (t<0);*
ii) For low α *(<.22 if b=0), the requirement that (i,t) induce the merchant to continue to sell to cash customers is a binding constraint on the EPN; for high enough* α *(above approximately.18 if b=0) the IR constraint binds.*
iii) When the IR binds, the sum of card user and merchant charges, i+t, is the same as the EPN's optimal choice under surcharging ((1+b)/2). As α *increases, i falls and t rises.*

Proposition 4i) follows from Proposition 2i) for the case where rebates were not feasible. Figure 1 illustrates the case where the merchant's IR constraint does not bind at i_0, the EPN's optimal merchant fee conditional on $t=0$. Recall from Lemma 3 that, for fixed $i+t$, the EPN wishes to lower i in the absence of other constraints. Thus, a movement down and to the right along the line $i+t = i_0$ (i.e., a cut in t and an equal increase in i) raises EPN profit. Expression (3) yields the IR constraint. Given linear demand, this constraint is linear with slope steeper than -1. Point B in Figure 1 represents the intersection of the line $i+t$ with this manifold. The EPN's solution is, then, to move down and to the right from $(i_0,0)$ to B along the line $i+t = i_0$, then down the IR line until it reaches an EPN indifference curve that is tangent to the IR (point C in Figure 1). This point represents a lower total EPN charge, $i+t$, and a lower t compared to $(i_0, 0)$. If, instead, the IR constraint binds at $t=0$ (IR cuts the horizontal axis at $i < i_0$), then with rebates the EPN immediately moves down and to the right along the IR manifold to a point of tangency. In both cases, therefore, $i+t$ is lower with rebates than in the NSR equilibrium under no rebates.

Turning to Proposition 4ii), if the cash market is sufficiently small then the floor on t is not the merchant's IR constraint but the need to induce the merchant to continue serving cash users; consequently, even with rebates feasible, the EPN cannot always fully extract the

merchant's surplus.

When the cash market is large enough that EPN charges are determined by the IR constraint, the total charge $i+t$ under the NSR is the same as under surcharging and is independent of the size of the cash market, α, but the spread between i and t (which is irrelevant under surcharging) shrinks as α increases. (Proposition 4iii.) These results can be understood as follows. The merchant's choices as functions of *(i, t)* are given by

$$p=\frac{1}{2}+\frac{i-b-t}{2(1+\alpha)}, \quad q_c=\frac{1}{2}-\frac{i-t-b}{2(1+\alpha)}, \quad q_e=\frac{1}{2}-\frac{i+t-b+2\alpha t}{2(1+\alpha)} \tag{4}$$

Note that q_e is decreasing in i and t, but the effect of t is stronger. (By contrast, i and t affect q_c symmetrically, since both operate on q_c only indirectly via the merchant's price p.) The EPN's equilibrium fees are then (see Appendix),

$$t^*=-\frac{1+b}{4(\alpha+\sqrt{\alpha}\sqrt{1+\alpha})}-\frac{b}{2}, \qquad i^*=\frac{1+b}{2}-t^*$$

The total charge when the merchant IR binds, t^*+i^*, is therefore $(1+b)/2$, the same as under surcharging, but lower than under the NSR with no rebates. Under the NSR with rebates, card transactions increase more if the total EPN fee is cut through rebates than through cutting i (see fn.18). The EPN prefers to grant rebates and reduce the total charge as needed to satisfy the merchant's IR because it gains enough from the increased transactions. As the cash market increases, the EPN continues meeting the IR with the same total charge (as opposed to cutting it under no rebates) but reducing the spread between i and t: t^* rises with α (smaller rebates) while i^* falls. The merchant benefits from this reduced spread because it gains the option of maintaining the same margin $p-i$ on cards but at a price p closer to the cash market optimum.

The next Proposition describes the effects of the NSR with rebates on quantities and welfare, when the cash market is large enough that EPN charges are determined by the merchant IR constraint ($\alpha > 0.22$ if $b=0$). Define the change in total surplus when rebates are allowed (ΔTS^R) to be total surplus under the NSR with rebates minus total surplus under no NSR. The change in aggregate consumer surplus (ΔCS^R) is defined analogously.

Proposition 5 (Quantities and Welfare): *Assume linear demand. Suppose an NSR is imposed and rebates are feasible. For* α *such that the merchant IR binds, compared to the equilibrium with no NSR:*

i) Cash users' transactions and consumer surplus are lower;

ii) Card users' transactions and consumer surplus are higher;

iii) Aggregate transactions $(q_e+\alpha\, q_c)$ *are unchanged;*

iv) ΔTS^R *rises in* α*. For b=0, it is positive if and only if* $\alpha > 1/3$*;*

v) ΔCS^R *falls in* α*. For b=0, it is negative if and only if* $\alpha > 1/3$*;*

vi) ΔTS^R *and* ΔCS^R *rise in b.*

Parts i)-iii) of Proposition 5 follow because the EPN's total charge $i+t$ is equal under the two regimes. Under surcharging, equilibrium quantities are invariant to how $i+t$ is divided between i and t, in particular, the same quantities would arise if one set (i^*, t^*) — the values that are optimal under the NSR. Imposing the NSR while charging (i^*, t^*), however, constrains the merchant's retail pricing, causing cash transactions to fall and card transactions to rise; total transactions remain the same because of the linearity of demand.

Now consider why ΔTS^R increases with the size of the cash market (Proposition 5iv). As α increases, the total quantity of transactions rises under both regimes but remains equal. Thus, the behavior of ΔTS^R hinges on the change in allocation of transactions between cash and card users. The efficient per-capita levels are 1 for cash and $1+b$ for cards. With surcharging, the cash quantity is $1/2$ and the card quantity is $(1+b)/4$, both independent of α. Under the NSR, as α increases both quantities rise towards $1/2$. As α increases, therefore, the allocation improves only under the NSR, so ΔTS^R rises in α. For $b=0$, ΔTS^R is positive if and only if $\alpha > 1/3$. The case of positive merchant benefit from card use, $b>0$, is discussed shortly

By contrast, the change in aggregate consumer surplus when moving to the NSR with rebates declines in α (Proposition 5v). The narrowing of the gap between per-capita cash and card quantities as α increases under the NSR — but not under surcharging — is harmful to overall consumer surplus, by the 'trapezoids argument' (see the discussion of Proposition 3ii)c)). For $b=0$, the NSR with rebates reduces overall consumer surplus if and only if $\alpha > 1/3$.

Moving from surcharging to the NSR with rebates, therefore, causes opposite changes in total surplus and overall consumer surplus if $b=0$: when $\Delta TS^R>0$ ($\alpha > 1/3$), $\Delta CS^R<0$. Thus, the increase in EPN profit always comes at least partly at the expense of consumers and the merchant

(recall that the merchant loses for all α). The case $b=0$, however, presents an overly negative picture of the NSR. When there are gross merchant benefits from card use, the NSR with rebates can increase both total surplus and overall consumer surplus. Since both ΔTS^R and ΔCS^R are increasing in b (Proposition 5vi) and both equal 0 at $\alpha = 1/3$, for $b>0$ there will be an interval around $\alpha=1/3$ in which ΔTS^R and $\Delta CS^R > 0$. These results are illustrated in Figure 2.

The intuition for why ΔTS^R and ΔCS^R are increasing in b is as follows, starting with total surplus. Since total transactions are equal under surcharging and under the NSR with rebates, the differential effect of b under these regimes works via its effect on the mix of transactions. The efficient per-capita cash and card quantities are $1+b$ and 1. With surcharging, quantities are $1/2$ and $(1+b)/4$, hence $q_e - q_c = (b-1)/4$. Recalling that $b<1$, by **A3)**, the card quantity is always lower than the cash quantity under surcharging, and the gap closes at the rate $\Delta b/4$. Under the NSR and rebates, q_e is higher than with surcharging, and any such gap is socially more valuable the larger is b. Moreover, as b increases the gap between card and cash quantities rises faster under the NSR than under surcharging and the gap under the NSR never exceeds the efficient gap, b. Thus, an increase in b magnifies the allocation advantage of the NSR. Finally, since the gap $| q_e - q_c |$ rises with b under the NSR but falls under surcharging, while total quantity is the same under both regimes, it follows that ΔCS^R increases in b (trapezoids argument).

Interestingly, the favorable impact of b on the NSR's effect on total surplus is reversed when rebates are not feasible (Proposition 3ii)c). With the NSR and no rebates, the EPN responds to an increase in b (in the range of α where the merchant's IR binds) by raising its charge to the merchant so as to leave the net marginal cost of card transactions, $i+t-b$, unchanged (Proposition 2iii), and therefore quantities unchanged.[27] Under surcharging, an increase in b leads the EPN to permit a reduction in $i+t-b$, and therefore an increase in card transactions. Thus, an increase in b raises card transactions under surcharging but not under the NSR with no rebates. Under the NSR with rebates, however, the EPN responds to an increase in b by allowing card transactions to rise *faster* than under no NSR, causing ΔTS^R to increase in b.

Proposition 6 draws on previous results to compare the outcomes under the NSR if rebates are or are not feasible. For simplicity, we focus on the case when the cash market is large

---

[27] Under the NSR and no rebates, when themerchant's IR binds, per capita transactions Q are determined by (3), whose right hand side is independent of b — since the merchant's outside option of serving only cash customers is independent of b, so too is the profit, and thus quantity, Q, that the EPN must leave to the merchant.

enough that the merchant's IR determines EPN pricing with or without rebates.

Proposition 6 (Rebates vs. No Rebates): *Assume linear demand. Suppose the NSR is imposed and rebates are feasible. Compared to the outcome under the NSR when rebates are not feasible:*
i) Card users' consumer surplus is higher, cash users' consumer surplus is lower, and aggregate consumer surplus is higher with rebates.
ii) For relative sizes of the cash market α *that make the merchant's IR constraint bind in both cases, total transactions and total surplus are higher with rebates.*

The superiority of rebates for total output and overall welfare (Proposition 6ii) reflects the ability of rebates under the NSR (and only then) to more effectively reduce double marginalization than by relying just on cutting the merchant fee. Recall that with the NSR an increase in i of Δ and an equal cut in t would lower the net price to card users, because the merchant would raise its uniform price by less than Δ. (In (4), $\Delta p = (\Delta i - \Delta t)/2(1+\alpha) = 1/((1+\alpha)$.) In fact, the price to card users is even lower under rebates, because when cutting t below 0, the EPN raises its merchant fee by less than the rebate amount. The lower aggregate EPN charge $i+t$ under rebates implies that total transactions increase. Aggregate consumer surplus therefore rises with rebates because total quantity is higher and per capita transactions are unequal under rebates but equal with no rebates (trapezoids argument). Total surplus is therefore also higher with rebates: overall consumer surplus is higher, the EPN's profit is higher (by revealed preference), and the merchant's profit is the same (for values of α that make the IR bind under rebates or no rebates, the merchant loses the entire surplus from dealing with the EPN in either case).

Cash users, however, lose from the NSR even with no rebates to card users (Proposition 3), and lose further if such rebates are feasible. Granting rebates increases the inverse demand of card users, prompting the merchant to raise its retail price.[28] In addition, when the EPN cuts t below 0 it also raises i somewhat, putting further upward pressure on the merchant's price (see (4) where p increases in i and decreases in t, while q_c does the reverse).

[28] Gerstner and Hess (1991) cite empirical evidence that retailers indeed raise their prices in response to manufacturers' granting of rebates to consumers.

6. Competitive Card Issuers

To this point, our analysis applies most directly to the case of proprietary networks where the EPN is a single card issuer. Alternatively, it describes outcomes when despite multiple card issuers, the issuing industry behaves as if were maximizing issuing banks' joint profits. How do the results change if the EPN is an association of *competitive* issuing banks? In this scenario, member banks issue the cards, and they, rather than the network, set most of the terms to cardholders, including prices (annual fee, interest rate, rebates). This section explores the effects of an NSR when the EPN is unable to control t.

A sequential/simultaneous game emerges. First, through their partnership with the EPN, banks set the merchant discount fee i and commit to it. Merchants continue to set prices taking i as given but recognizing that t is determined through competition for card users by issuing banks. If bank member W of the EPN is one of m banks charging the lowest card user fee, it obtains sales of $q_W = x/m$, where x is derived from equation (1) and is given by $x=1/2-(i+t-b)/2$. If the fee of bank W is not among the lowest, q_W is zero. That is, taking i as given, banks compete as Bertrand price setters to cardholders and each of the banks that charge the lowest fee t obtains $1/m$ of total transactions, where the latter quantity x is determined by the equality of the merchant's marginal revenue function from card transactions with its net marginal cost. Suppose that banks can only set fees in discrete units, $\in \approx 0$. By the standard Bertrand logic, the equilibrium t_W satisfies $t_W = -i + \in$. Card issuers compete away (virtually) all their rents by offering rebates that are close to the interchange fee.

As before, the constraints on (i,t) are to ensure the merchant continues serving cash users, and continues participation with the EPN (IR). In both cases, equilibrium quantities are obtained by substituting $t_W \approx -i$ into expressions (4) that show the merchant's quantities as functions of i and t. Since the quantities under no NSR are $q_c = 1/2$, $q_e = (1+b)/2$, the changes are

$$\Delta q_c=(2t+b)/(2(1+\alpha))<0, \qquad \Delta q_e=-\alpha(2t+b)/(2(1+\alpha))>0.$$

The inequalities follow since the NSR binds on the merchant only if $t < -b/2$[29]. The changes in equilibrium quantities imply that with competitive issuers total transactions under the NSR with rebates is the same as under no NSR. For $b=0$, the per-capita card quantity exceeds the cash

[29] The Bertrand assumption implies $t \approx -i$. Linear demand implies that, under surcharging, the merchant's card price is $(1-b-2t)/2$. This exceeds the cash price $(1/2)$: that is, the NSR binds only if $t<-b/2$.

quantity under no NSR (because, with competitive issuers, the markup is only at the merchant level), but exceeds it under the NSR with rebates.

As long as the EPN's issuing banks enjoy *some* profits from transactions ($\epsilon > 0$), the EPN will wish to generate the largest possible quantity of such transactions. Since card transactions are decreasing in t, the EPN will fix a high i, inducing its competing issuers to offer large negative values of t (large rebates). Proposition 7 summarizes the effects of this incentive on equilibrium quantities under the NSR and competitive issuers. [30]

Proposition 7: *Assume linear demand and b = 0. With perfectly competitive issuers, in the equilibrium under the NSR:*
i) If $\alpha<1$, the EPN sets i until merchants are just indifferent between selling to cash customers or not; if $\alpha>1$, the merchant's IR constraint binds;
ii) Cash transactions are lower than with no NSR, card transactions are higher, but total transactions are the same;
iii) For all values of α, overall consumer surplus is higher than with no NSR but merchant profit and total surplus are lower;
iv) In the limit as the mass of cash users becomes large, the per-capita cash quantity approaches the single monopoly level and the per capita card quantity approaches the competitive level.

Proposition 7i) illustrates that, with competitive issuers, the constraint that the EPN ensures that the merchant continues to serve the cash market binds for a larger size of the cash market ($\alpha \leq 1$ rather than $\alpha \leq .22$). This is because the stronger tendency to offer rebates under competition among card issuers makes the option of pricing cash users entirely out of the market relatively more attractive to merchants. Total quantity remains the same as under no NSR (Proposition 7ii)) because demand is linear and the total EPN fee remains the same (zero). Card and cash quantities therefore move in opposite directions because only card users get rebates.

Given the same total quantity and $b=0$, total surplus must fall under the NSR with rebates, since per capita quantities of cash and card users are then different, while efficiency calls

---

[30] The theorem is shown for $b = 0$, however, given the continuity of the environment, quantity and welfare results will continue to hold for b small and positive. They may not hold for b large since, even with competitive issuers, there is then a significant bias away from cards under no NSR. (The efficient quantities are 1 for cash and $1+b$ for cards while the no NSR levels are $1/2$ and $(1+b)/2$, so only the card underprovision rises with b.)

for equal levels as occurs with competitive issuers and no NSR. This divergence of quantities only with the NSR implies, however, that overall consumer surplus rises (trapezoids argument).

Result 7iv) shows that as the cash market becomes large relative to cards, the NSR in conjunction with competitive rebates by card issuers succeed in eliminating the distortion in the pricing of card transactions due to the monopolist merchant. The merchant charges a uniformly high (monopoly) price to both card and cash users, but card users receive a rebate and therefore obtain a net price close to the competitive price. However, the net price to cash users is the (uniform) price charged by the merchant. When the cash market is large, the merchant's price is driven by the cash market and thus will approach the simple monopoly level.

7. Conclusion

The complex cycle that makes up a typical payment network offers a rich field for economic analysis, with prices playing important roles at every link of the cycle. Our principal model analyzed the No Surcharge Rule as an imperfect instrument of vertical control by a card payment network (EPN) facing a merchant in an environment of double marginalization, where the merchant also serves outside consumers—'cash' users. By requiring the merchant's card price to equal its cash price, the NSR leads the EPN to prefer a higher fee to the merchant and a lower fee to card users (whereas the EPN is indifferent to how it allocates its total fee when the merchant can set the card price independent of the cash price). Throughout, the NSR benefits the EPN but harms the merchant and cash users. Other welfare effects depend on the ratio of cash to card users, the merchant's benefit from card versus cash transactions, and whether rebates (negative fees) to card users are feasible.

If rebates are not feasible, the EPN charges card users zero but raises its merchant fee above its no-NSR total fee. This increase in total fee reduces total transactions and aggregate consumer surplus; with a sufficiently small cash market, even card users pay more under the NSR. Despite the fall in total quantity, overall welfare increases if (and only if) the cash market is large enough, because the rise in per capita card quantity combats the pre-NSR bias from double marginalization, at the cost of a relatively small distortion in *per capita* cash quantity.

If rebates are feasible the EPN grants them, benefiting itself and card users while harming cash users and (weakly) the merchant. However, the EPN's total fee is lower with rebates (as needed to maintain merchant participation), so total quantity is higher than under the NSR with

no rebates, as are aggregate consumer surplus and overall welfare. Relative to no NSR, the NSR with rebates leaves total quantity unchanged but reverses the gap between the card and cash quantities from negative to positive. Overall welfare rises if and only if the cash market is sufficiently large, because the per capita cash distortion then is small (though a larger cash market makes the NSR less favorable to total consumer surplus). A larger merchant benefit from card compared to cash transactions increases the pre-NSR distortion from double marginalization in card pricing and improves the effects of the NSR with rebates on overall welfare and total consumer surplus; interestingly, the reverse occurs if rebates are not feasible.

Our emphasis has been on the impact of limits on merchant pricing flexibility when there exists some power over price. To highlight this effect, we assumed monopoly pricing at both the merchant and EPN levels, but we conjecture that similar effects will arise whenever there remains a significant margin between price and marginal cost at both levels. We also analyzed a case where the EPN margin is almost zero, because the EPN's card issuing banks behave as Bertrand competitors. In that case, pre-NSR there is no significant bias against cards (if merchant benefit from cards is low), so by encouraging card transactions at the expense of cash the NSR with rebates reduces welfare, though it increases overall consumer surplus. Our analysis abstracted away from consumers' choice of the means of payment in order to focus on the impact on the level of transactions per consumer. Extensions of this research would include analyses of the effects of the 'rule' under a broader class of merchant market structures and with endogenous consumer choice of the means of payment (e.g., Rochet, 2003). Another direction will be to examine how the NSR influences the competition among rival payment networks both in pricing and in other practices such as the tying of multiple cards (e.g., Rochet and Tirole, 2003).

Appendix

Proof of Lemma 1: Let $x=x(k)$, $x'=x(k')$. By definition, $(V'(x)-k)x \geq (V'(x')-k)x'$ and $(V'(x)-k')x \leq (V'(x')-k')x'$ so $(k'-k)x \geq (k'-k)x'$.This implies $x \geq x'$. Now, suppose $x=x'$. The first order condition for $x(k)$ satisfies $xV''(x)+V'(x)-k=0$. For $k'>k$, then, $xV''(x)+V'(x)-k'<0$. ‖

Proof of Lemma 2: The derivative of the EPN's profit function is

$$d\Pi^e(x)/dx=(xV''(x)+V'(x))+x^2 V'''(x) +2x V''(x) +b.$$

At x_0, the term in parentheses is zero by the optimality of x_0 for the merchant in the cash market. Assumption **A3)** implies the remaining term is negative. The concavity of the EPN profit function implies the right side declines as x rises. Therefore, EPN profits are declining in x for $x \geq x_0$. ‖

Proof of Lemma 3: (i) By Lemma 1, $x(k-b) < x_0$ and therefore, $V'(x(k-b))>V'(x_0)$. Since $t^* \equiv V'(x(k-b))-V'(x_0)$, for all t,i such that $i+t=k$, $x(k-b)$ remains constant and $t<t^*$ implies $V'(x(k-b))-t=p_e^M>V'(x_0)=p_c^M$.

(ii) Consider a choice of (q_e, q_c) that solves the merchant's profit maximization problem with an NSR. Suppose that $q_e < x(k-b)$. The pair $(x(k-b),q_c)$ is also feasible for the merchant since $V'(q_c)+t \geq V'(q_e)$ implies $V'(q_c)+t \geq V'(x(k-b))$ by the concavity of $V(\bullet)$. But the choice of (q_e, q_c) over $(x(k-b),q_c)$ then implies that

$$q_e (V'(q_e)-k+b) \geq x(k-b)(V'(x(k-b))-k+b)$$

which violates the definition of $x(k-b)$. A similar proof shows $q_c \leq x_0$. Now suppose $q_e=x(k-b)$. The merchant's first order condition with respect to q_e under the NSR constraint is $q_eV''(q_e)+V'(q_e)-i-t+b-\lambda V''(q_e)$ where $\lambda > 0$ is the multiplier on the constraint imposed by the NSR. Evaluating this expression at $x(k-b)$ yields $-\lambda V''(x(k-b))>0$ since the first terms are the merchant's first order condition with no NSR and equal zero at $x(k-b)$. Therefore, merchant profits are strictly increasing in q_e at $q_e=x(k-b)$.

(iii) From **A1)**, $f(p)$ is the demand curve of cash users with $pf(p)$ concave. The demand curve of card users is $f(p+t)$. Let $i+t = k$ and let p denote the optimal (uniform) price charged by the merchant under an NSR when the card user fee is t (so $i = k-t$). Similarly, let p' denote the optimal uniform price charged by the merchant when the card user fee is $t' < t$. Finally, for convenience, set $t-t' \equiv \Delta > 0$. By definition of p, charging a price p under the fee profile, $(k-t,t)$ yields higher merchant profits than charging a price $p'-\Delta$. Note that this second price implies a

net price to card users of $p'+t'$. Thus,

$$\alpha pf(p)+(p+t-(k-b))f(p+t) \geq \alpha(p'-\Delta)f(p'-\Delta)+(p'+t'-(k-b))f(p'+t').$$

Similarly, under the fee profile, $(k-t',t')$, p' raises more profits than charging a price $p+\Delta$.

$$\alpha p'f(p')+(p'+t'-(k-b))f(p'+t') \geq \alpha(p+\Delta)f(p+\Delta)+(p+t-(k-b))f(p+t).$$

Adding the two inequalities and eliminating the common terms which denote revenues in the card market and dividing by α, yields

$$p\,f(p)-(p+\Delta)\,f(p+\Delta) \geq (p'-\Delta)f(p'-\Delta)-p'f(p').$$

Recall that p and p' are higher than the price which maximizes $p\,f(p)$. Suppose that $p'+t'>p+t$. This implies $p'-\Delta>p$. But this violates the assumption of concavity of $p\,f(p)$ since the slope of the revenue function must become steeper as we move further to the right of the maximum point. ‖

Proof of Proposition 2: i) When the IR constraint does not bind, the result follows from Lemma 3iii). Now suppose the IR binds and consider *(t,i)* space. (For the purposes of this argument, *(t,i)* space is more convenient than *(i,t)* space shown in Figure 1.) At $t = 0$, and i such that the merchant IR curve binds, the slope of the EPN level set is steeper than the slope of the merchant IR curve. This implies that this point is a local maximium. Under an NSR, the Lagrangian representing the merchant's profit maximization problem is

$$L(q_c\ q_e,\lambda;i,t)=\alpha\, q_c\, V'(q_c)+ q_e\,(V'(q_e)-i-t+b)+\lambda(V'(q_c)+ t - V'(q_e)).$$

The first order conditions from this problem satisfy

$$(\lambda+ \alpha\, q_c\,)V''(q_c) +\alpha\, V'(q_c) = 0, \tag{1A}$$

$$(-\lambda+ q_e)V''(q_e) + V'(q_e) -i-t +b = 0, \tag{2A}$$

$$V'(q_c)+ t - V'(q_e) = 0. \tag{3A}$$

Totally differentiate this system and evaluate it at $t = 0$, so $q_c = q_e \equiv Q$: so Q repesents the common per capita quantity, given i (and $t=0$). Define $V''(Q) \equiv V''$, and $V'''(Q) \equiv V'''$. This yields the system

$$\begin{bmatrix} V'' & V'''(\lambda+\alpha Q)+2\alpha Q V'' & 0 \\ -V'' & 0 & V'''(\lambda+\alpha Q)+2\alpha Q V'' \\ 0 & V'' & -V'' \end{bmatrix} \begin{bmatrix} d\lambda \\ dq_c \\ dq_e \end{bmatrix} = \begin{bmatrix} 0 \\ di+dt \\ -dt \end{bmatrix}$$

Using (1A) to substitute in for λ, we can solve this system to obtain (at $t = 0$)

$$\frac{\partial q_e}{\partial t}=\frac{1+2\alpha-\alpha V'''\dfrac{V'}{V''}}{(1+\alpha)(QV'''+2V'')} \;\le\; \frac{\partial q_e}{\partial i}=\frac{1}{(1+\alpha)(QV'''+2V'')}<0.$$

The first inequality follows from the concavity of the merchant revenue function in price and the second from the concavity of the revenue function in quantity (**A1)**. Equations (1A) and (2A) imply that $i = b - \lambda\,(1+\alpha)\,V''/\alpha$. This yields

$$i\left(\frac{\partial q_e}{\partial i}-\frac{\partial q_e}{\partial t}\right) \;=\; b\left(\frac{\partial q_e}{\partial i}-\frac{\partial q_e}{\partial t}\right)+\lambda\frac{2V''-V'''\dfrac{V'}{V''}}{QV'''+2V''} \;\ge\lambda\frac{2V''-V'''\dfrac{V'}{V''}}{QV'''+2V''}$$

and

$$i\frac{\partial q_e}{\partial i}+Q \;=\; b\frac{\partial q_e}{\partial i}-\frac{\lambda V''}{\alpha(QV'''+2V'')}+Q \;\le\frac{V'+QV''}{(QV'''+2V'')}+Q.$$

The inequality follows because q_e is decreasing in i and from the substitution for λ . Combining these results yields

$$\frac{i\left(\dfrac{\partial q_e}{\partial i}-\dfrac{\partial q_e}{\partial t}\right)}{i\dfrac{\partial q_e}{\partial i}+Q}\ge\frac{\lambda(2V''-V'''\dfrac{V'}{V''})}{V'+3QV''+Q^2V'''}.$$

Now consider the level sets of the merchant and the EPN in *(t,i)* space. The slopes at $t = 0$ are given by

$$\frac{di^e}{dt}_{t=0}=-\left(1-\frac{i\left(\dfrac{\partial q_e}{\partial i}-\dfrac{\partial q_e}{\partial t}\right)}{i\dfrac{\partial q_e}{\partial i}+Q}\right),\qquad \frac{di^M}{dt}_{t=0}=-\left(1-\frac{\lambda}{Q}\right).$$

Subtracting the second from the first yields, after substituting the inequality from above,

$$\frac{di^e}{dt}_{t=0} - \frac{di^M}{dt}_{t=0} \geq \left(\frac{-\lambda}{V''Q}\right)\left(\frac{QV''+V'}{V'+3QV''+Q^2V'''}\right)\left(V''+V'''Q\right) \geq 0.$$

The first term is positive because λ is non-negative and V is strictly concave. The numerator in the second term is positive because $Q \leq x_0$ and $xV''(x) + V'(x)$ is decreasing. The denominator is the marginal revenue curve of the EPN and is non-positive because if the IR constraint is binding, i is less than what it would choose (at $t = 0$) if the IR did not bind. The final term is negative by (1A). Since the slope of the EPN level set exceeds the slope of the merchant's IR curve at $t = 0$, this point represents a local maximum. By Assumption **A4)**, this is also a global maximum. Thus 2i) follows.

ii) Note that at $t = 0$, equations (1A) and (2A) imply that

$$(1+\alpha)(Q^2 V'' + QV') = (i-b)\, Q$$

(Note that this implies that the merchant's choice of Q is a strictly decreasing function of $(i-b)$.) Substituting in for $(i-b)\, Q$ the IR constraint is equivalent to

$$(1+\alpha)\, QV'(Q) - (i-b)Q = -(1+\alpha)\, Q^2 V''(Q) \geq \alpha\, x_0\, V'(x_0)$$

or (Equation (3) in the text)

$$-Q^2 V''(Q) \geq \alpha\, x_0\, V'(x_0)/(1+\alpha)\,. \qquad (4A)$$

The right side is increasing in α. Concavity of the merchant revenue function in quantity, **A1),** implies the left side is increasing in Q. For low α, the constraint does not bind when the EPN selects its globally optimal Q at $(i_0, 0)$. As α rises, the constraint binds and the EPN must offer a successively higher Q (lower i) in order to induce the merchant to participate.

iii) From (1A) and (2A), the merchant choice of Q is strictly increasing in i-b, so, holding b fixed, 2ii) implies i decreasing in α. When the IR binds, Q is determined by (4A) which, in turn, determines i-b.

iv) *If the IR does not bnd, then the EPN optimal choice of i is* $(1+\alpha+b)/2$ *which is increasing in* α. *If the IR binds, then the optimal choice of i is determined solely by the merchant's IR constraint (at t=0) and is given by*

$$1+\alpha-\sqrt{\alpha+\alpha^2}+b$$

This is decreasing in α *for all* α.

‖

Proof of Proposition 3: i) Proposition 2 yields the optimal solution $t = 0$ which implies that per capita cash and card purchases are the same. This gives the first order condition of the merchant, $i = b+(1+\alpha)(V'(x)+xV''(x))$. Define $q_\alpha = \text{argmax}_x\, x(b+(1+\alpha)(V'(x)+xV''(x)))$ to be the quantity of card-user transactions which maximizes EPN profits with the NSR. Assumption **A2)** implies this is unique. Note that q_0 maximizes profits with no NSR. If $b=0$, then the definition indicates that $q_0 = \text{argmax}_x\, (1+\alpha)\, x((V'(x)+xV''(x)))$ and so q_0 also solves the EPN's problem with the NSR. Thus, card transactions are unchanged with the NSR (and $b=0$) but cash transactions are lower (since they exceeded card transactions without the NSR). Now consider $b > 0$. By definition,

$$q_0 (b+V'(q_0)+q_0 V''(q_0)) \geq q_\alpha(b+V'(q_\alpha)+q_\alpha V''(q_\alpha))$$

and

$$q_0 (b+(1+\alpha)(V'(q_0)+q_0 V''(q_0))) \leq q_\alpha(b+(1+\alpha)(V'(q_\alpha)+q_\alpha V''(q_\alpha))).$$

Subtract the two inequalities and divide by $-\alpha$ to get

$$q_0(V'(q_0)+q_0 V''(q_0)) \leq q_\alpha(V'(q_\alpha)+q_\alpha V''(q_\alpha)).$$

Suppose that $q_\alpha > q_0$. Then $b\, q_0 < b\, q_\alpha$. This implies

$$q_0(b+V'(q_0)+q_0 V''(q_0)) < q_\alpha(b+V'(q_\alpha)+q_\alpha V''(q_\alpha))$$

which violates the definition of q_0. The EPN first order conditions with the NSR, evaluated at q_0, indicates that EPN profits are strictly declining in quantity at that point:

$$\frac{\partial \pi^e(x)}{\partial x}_{|x=q_0} = b(1-(1+\alpha)) = -\alpha b < 0$$

so $q_\alpha < q_0$ given $b>0$. Thus, card transactions are lower with the NSR for $b>0$ and so, too, are cash transactions.

ii) a) - b) The limit of the right side of (4A) as α becomes large is $x_0V'(x_0)$ so Q must approach x_0 Before reaching the limit, though, $Q < x_0$ so cash users' purchases and surplus fall with the NSR. Lemma 2 implies that eventually cardholder purchases and surplus are higher with the NSR.

c) With an NSR and linear demand, merchant profit is $(1+\alpha)Q^2$ so (3) can be written $Q^2 \geq \alpha/(4(1+\alpha))$. With no NSR, the total quantity of transactions is $(1+b)/4+\alpha/2$. For $\alpha > \alpha^*$, the IR constraint binds and determines $Q=(\alpha/(4(1+\alpha))^{.5}$. The NSR thus raises total quantity if and only if,

$$(1+\alpha)\sqrt{\frac{\alpha}{4(1+\alpha)}}=\frac{\sqrt{1+\alpha}\sqrt{\alpha}}{2}\geq\frac{1+b}{4}+\frac{\alpha}{2}$$

$$\alpha+\alpha^2\geq\frac{1+b^2}{4}+(1+b)\alpha+\alpha^2$$

This is impossible so total quantity falls. The limit as α goes to infinity of the difference in total quantity is

$$\lim_{\alpha\to\infty}(1+\alpha)Q-\alpha/2-(1+b)/4=\frac{1}{2}\lim_{\alpha\to\infty}\left(\sqrt{(1+\alpha)\alpha}-\alpha\right)-\frac{1+b}{4}$$

$$=\frac{1}{2}\lim_{\alpha\to\infty}\left(\frac{\sqrt{\frac{1+\alpha}{\alpha}}-1}{1/\alpha}\right)-\frac{1+b}{4}$$

Applying L'Hopital's Rule to the limit (the term in the limit is *1/2)* yields a value for the difference in total quantity as α goes to infinity is *-b/4.*

For any two pairs of per capita transactions, $(q_c,q_e),(Q_c,Q_e)$, and defining,

$\Delta Q_c = Q_c - q_c, \Delta Q_e = Q_e - q_e, \Delta Q_T = \alpha\Delta Q_c + \Delta Q_e,$

the change in total surplus when moving from the first outcome to the second is

$$\Delta TS = (1-.5(Q_c+q_c))\Delta Q_T + (b-((Q_e-Q_c)+(q_e-q_c))/2)\Delta Q_e \ . \qquad (5A)$$

Thus, let $q_c = 1/2$, $q_e=(1+b)/4$ be the per capita transactions when surcharging is allowed and $Q=Q_e=Q_c$ the (common) per capita quantity under the NSR without rebates. The IR constraint yields $Q^2=.25\,\alpha/(1+\alpha)$, so the limit as α goes to infinity of Q is ½. Thus, as α goes to infinity, $\Delta Q_c = 0$, $\Delta Q_e = (1-b)/4$, $\Delta Q_T = -b/4$ and we have

$$lim_{\alpha\to\infty}\ \Delta TS^{NR} = (1+b)^2/32 - b^2/4.$$

So the limit is decreasing in *b.* Furthermore, using the fact that Q and q_c are independent of *b* and $\partial q_e/\partial b=1/4$, we have

$$\partial\Delta TS^{NR}/\partial b=-\partial q_e/\partial b+q_e\,\partial q_e/\partial b+Q-q_e-b\,\partial q_e/\partial b.$$

Therefore,

$$\partial\{\partial\Delta TS^{NR}/\partial b\}/\partial\alpha=\partial\, Q/\partial\alpha>0.$$

Direct computation shows that ΔTS^{NR} is increasing in α for $b=0$. Thus, it is increasing in α for all $b>0$. Since, for α large enough, $b' > b$ implies $\Delta TS^{NR}(\alpha,b')<\Delta TS^{NR}(\alpha,b)$ (the limit is decreasing

in *b),* and since $\partial \Delta TS^{NR}/\partial\alpha$ is increasing in *b,* we have $\Delta TS^{NR}(\alpha,b')<\Delta TS^{NR}(\alpha,b)$ for all α. (Computations show that total surplus exceeds total surplus with no NSR at $\alpha > 1.53$ for $b=0$.).

To see that the net effect on consumer surplus is negative, note that if total quantity had remained constant with card use rising and cash use falling, consumer surplus would have to fall since the loss in consumer surplus for cash users -- whose initial per capita quantity exceeded that of card users -- would necessarily be larger than the gain enjoyed by card users. Since, in fact, total quantity actually falls, the consumer surplus effects are worse. Formally, the change in consumer surplus is

$$\Delta CS^{NR} = .5(Q_c+q_c)\Delta Q_T + .5((Q_e-Q_c)+(q_e-q_c))\Delta Q_e \,, \qquad (6A)$$

and in this case, $\Delta Q_T <0$, $Q_e-Q_c=0$, and $(q_e-q_c)\Delta Q_e <0$. ‖

Proof of Proposition 4:i) If $\alpha<\alpha^*$, then the IR does not bind under no rebates and, by Lemma 3iii), the EPN increases profits by holding *i* fixed and lowering *t.* If $\alpha \geq \alpha^*$, from the Proof of Proposition 2i) the slope of the EPN indifference curve in *(i,t)* space is steeper than the slope of the merchant's IR curve at the constrained optimal solution, $t=0$. Thus, again, EPN profits are strictly higher as *(i,t)* is varied by lowering *t* below zero and raising *i* so as to stay on the IR curve.

ii) With $t<0$, the merchant's demand curve for card transactions is strictly above the (per capita) cash demand curve. With sufficiently high rebates, the monopoly price from serving the card market exceeds the choke price for the cash market. In this case, the merchant's profit function has two local maxima. If the merchant serves only the card market, the per capita card transaction is the same as with surcharging allowed but cash transactions are now zero. If the merchant prices to serve both markets, the solution is

$$p=\frac{1+\alpha+i-b-t}{2(1+\alpha)}, \quad q_c=\frac{1+\alpha-i+b+t}{2(1+\alpha)}, \quad q_e=\frac{1+\alpha-i+b-t(1+2\alpha)}{2(1+\alpha)} \qquad \textbf{(7A)}$$

The merchant chooses to serve both markets and thus selects prices and quantities as in (7A) if and only if

$$i-t-b\leq\sqrt{1+\alpha} \qquad \textbf{(8A)}$$

Using the values for $p,\ q_c, q_e$ from (7A) in (4A) gives the values of *(i,t)* for the EPN when the IR

binds. Maximizing with respect to *(i,t)* yields

$$t^* = -\frac{1+b}{4(\alpha+\sqrt{\alpha}\sqrt{1+\alpha})} - \frac{b}{2}, \qquad i^* = \frac{1+b}{2} - t^* \tag{9A}$$

Equation (9A) implies that (8A) is violated as α gets small.

iii) Follows from Equation (9A). ‖

Proof of Proposition 5:i)-iii) Follows from (7A) and noting that under surcharging, per capita cash quantity is *1/2* and card quantity is *(1+b)/4*.

iv)-v): Total surplus is affected by total quantity and by the differences in quantities. The NSR and linear demand imply $q_c^{NSR} = q_e^{NSR} + t$. Utilizing Equation (5A) for the change in total surplus moving from surcharging to an NSR with rebates, along with the fact that total quantity is unchanged yields

$$\Delta TS^{NSR} = (b-(-t-(1-b)/4)/2)\Delta Q_e = (7b+1+4t)\Delta Q_e/8.$$

The change in total surplus is positive if and only if $(7b+1+4t)>0$. It is increasing in α if $(7b+1+4t)>0$ since ΔQ_e and t are increasing in α. Note that if $b=0$, then $TS^{NSR} - TS^{SUR} = 0$ at $\alpha = 1/3$. (At that point, $t=-1/4$ and q_e^{NSR} exceeds q_c^{NSR} by exactly the same amount that q_c^{SUR} exceeds q_e^{SUR}.)

Given constant total quantity and linear demand, aggregate consumer surplus depends on the split between the types of consumers. Equation (6A) yields

$$\Delta CS^R = -(q_e^{NSR} - q_e^{SUR})[(q_c^{NSR} - q_e^{NSR}) + (q_c^{SUR} - q_e^{SUR})].$$

Direct computation yields that ΔCS^R is increasing in α for all b. Since ΔTS^R depends on α as ΔCS^R depends on -α, we also have ΔTS^R falls in α for α such that $(7b+1+4t)<0$.

vi) Direct computation shows that ΔTS^R and ΔCS^R rise in b. ‖

Proof of Proposition 6:i) Equation (7A) shows that q_c falls as $i-t$ rises and q_e rises if $i+t$ and t fall. Propositions 2 and 4 reveal that compared to the NSR with no rebates, when rebates are feasible, $i+t$ and t are lower and $i-t$ is higher. If the IR binds, then Propositions 3 and 4 imply total quantity is higher under the NSR with rebates and since, with no rebates, per capita quantities are always identical, $(q_e=q_c)$ and with rebates, $Q_e > Q_c$, Equation (6A) implies total consumer surplus must rise.

ii) Suppose that the IR binds at the optimal solution with $t=0$. The optimal solution with the $t \geq 0$ relaxed is at a point downward and to the right of this point. Part i) implies total consumer surplus rises. Revealed preference implies that EPN profits rise and, since we remain on the merchant's IR curve, merchant profits stay the same. Thus, total surplus rises when the $t \geq 0$ constraint is relaxed from a point at which the IR constraint binds.
iii) Shown by computation.

Proof of Proposition 7: i): Solving the merchant participation constraint simultaneously with the constraint $t=-i$, yields

$$i^{comp(IR)} \leq \frac{1}{2}\frac{\sqrt{1+\alpha}}{\sqrt{\alpha}}$$

The constraint that the merchant continue to be willing to serve the cash market is $t>i-(1+\alpha)^{.5}/2$. Using $t=-i$, this yields a value

$$i^{comp(IC)} \leq \frac{1}{2}\sqrt{1+\alpha}$$

The lowest value for i^{comp} is the binding constraint. The second one is lower than the first if and only if $\alpha<1$.
ii),iii): Use the quantity equations from Equation (7A) for per capita purchases and $t=-i$ to get $\alpha q_c=\alpha(.5-i/(1+\alpha))$ and $q_e=(.5+\alpha\, i/(1+\alpha))$. Summing the two yields total quantity $(1+\alpha)/2$ which is independent of i and is equal to the total quantity of purchases with competitive issuers and no NSR.. Conditional on total quantity remaining constant, social surplus is maximized when the cash and non-cash quantities are the same. Any value of t strictly less than zero along with the NSR, violates this condition, so social surplus must fall. Consumer surplus rises because, holding total quantity fixed, the loss to cash consumers from the higher price is more than compensated by the gain to EPN consumers from the lower price. Using $q_e=(.5+\alpha\, i/(1+\alpha))$ and letting α grow large yields i approaches $1/2$, EPN quantity approaches 1 and cash quantity approaches $1/2$. ‖

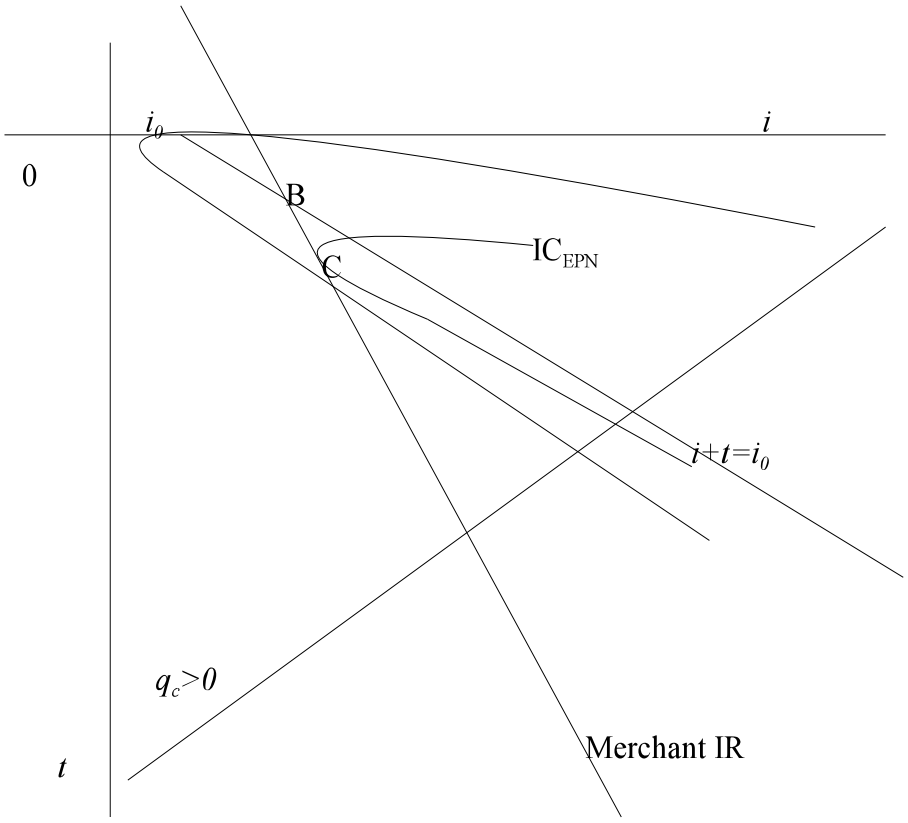

Figure 1

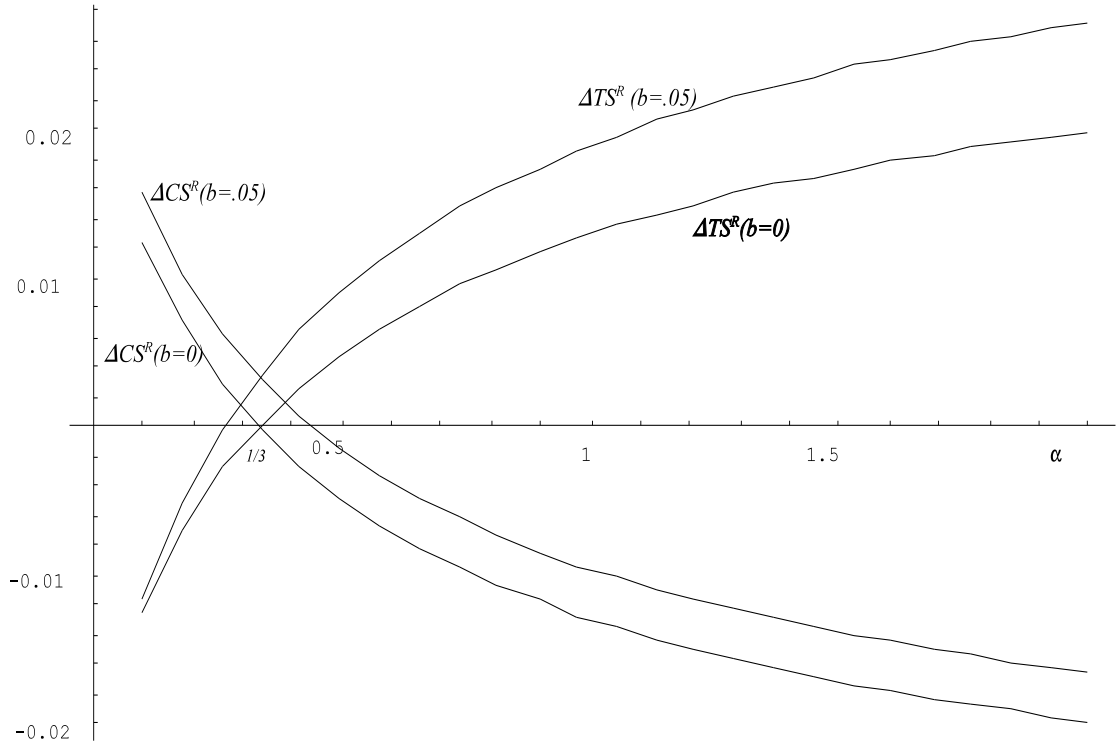

Figure 2

References

Baxter, William (1983). "Bank interchange of transactional paper: Legal and economic perspectives." *Journal of Law and Economics 26* (October): 541-588.

Carleton, Dennis and Alan Frankel (1995a). "The antitrust economics of credit card networks." *Antitrust Law Journal 63 (2)*: 643-668.

Carleton, Dennis and Alan Frankel (1995). "The antitrust economics of credit card networks: Reply to Evans and Schmalensee Comment." *Antitrust Law Journal 63 (3)*: 903-915.

Chain Store Age, *Fourth Annual Survey of Retail Credit Trends,* January 1994, section 2.

Chakravorti, Sujit (2003). "Theory of credit card networks: A survey of the literature." *Review of Network Economics* , Vol. 2, Issue 2 (June): 50-68.

Chakravorti, Sujit and Williams Emmons (2001). "Who pays for credit cards". Federal Reserve Bank of Chicago. EPS-2001-1.

Chakravorti, Sujit and Alpah Shah (2001). "A study of the interrelated bilateral transactions in credit card networks". Federal Reserve Bank of Chicago. EPS-2001-1.

Evans, David and Richard Schmalensee (1995). "Economic aspects of payment card systems and antitrust policy toward joint ventures." *Antitrust Law Journal 63 (3)*: 861-901.

Evans, David and Richard Schmalensee (1999). *Paying with Plastic: The Digital Revolution in Buying and Borrowing.* MIT Press.

Faulkner and Gray (2000), *Card Industry Directory*, 2000 edition, Chicago.

Federal Reserve Board (2001). *Recent Changes in Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances.* http://www.federalreserve.gov/pubs/oss/oss2/2001/bull0103.pdf

Gans, J. and King, S. (2003). "The Neutrality of Interchange Fees in Payment Systems." *Topics in Economic Analysis & Policy:* Volume 3, Issue 1. Article 1.

Gerstner, Eitan, and James D. Hess (1991). "A theory of channel price promotions." *American Economic Review* 81 *(September)*: 872- 886.

Katz, Michael L. *Reserve Bank of Australia. Reform of Credit Card Schemes in Australia II: Commissioned Report*, www.rba.gov.au , August 2001.

Malueg, David (1992). "Direction of price changes in third-degree price discrimination:

Comment." Freeman School of Business, Tulane University, Working Paper 92-ECAN-95.

Nahata, Babu, Krzysztof Ostaszewsi, and P.K. Sahoo (1990). "Direction of price changes in third-degree price discrimination." *American Economic Review* 80 *(December)*: 1254-1258.

The Nilson Report, May, 2000.

Reserve Bank of Australia (2002). *Reform of Credit Card Schemes in Australia IV: Final Reforms and Regulation Impact Statement.* 27 August.

Rochet, J.C. (2003). "The theory of interchange fees A synthesis of recent contributions." *Review of Network Economics* , Vol. 2, Issue 2 (June): 97-124.

Rochet, Jean-Charles and Jean Tirole (2002). "Cooperation among competitors: Some economics of payment card associations," *Rand Journal of Economics*, Vol. 33, 549-570.

Rochet, Jean-Charles and Jean Tirole (2003). "Platform competition in two-sided markets." *Journal of the European Economic Association,* forthcoming.

Salop, Steven (1990), "Deregulating self-regulated shared ATM networks," *Economics of Innovation and New Technology* volume 1, numbers 1-2 (December): 85-96.

Schmalensee, Richard (2002). " Payment systems and interchange fees." *Journal of Industrial Economics*, Vol. 50, 103-122.

Tirole, Jean (1988). *The Theory of Industrial Organization.* MIT Press.

Wright, Julian (2000). "An Economic Analysis of a Card Payment Network." Attachment 2, *Credit Card Schemes in Australia.*. VISA International Service Association. January, 2001.