The No Surcharge Rule and Card User Rebates: Vertical Control by a Payment Network

MARIUS SCHWARTZ

Georgetown University

DANIEL R. VINCENT *

University of Maryland, College Park

Abstract

The No Surcharge Rule (NSR) prevents merchants from charging more to consumers who pay by card versus other means ("cash"). We consider a payment network facing local monopolist merchants that serve two consumer groups, card users and cash users. Unlike in prior work, transaction quantities are variable. The NSR raises network profit and harms cash users and merchants; overall welfare rises if and only if the ratio of cash to card users is sufficiently large. With the NSR, the network will grant rebates to card users whenever feasible. If rebates are not feasible, the NSR can harm even card users.

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 practice involves constraints on the

* Contact author. Department of Economics, University of Maryland, College Park, MD 20742. dvincent@umd.edu For helpful comments and suggestions, we would like to thank various seminar participants including Andrew Dick, Patrick Greenlee, Bob Hunt, David Malueg, Alex Raskovich, George Rozanski and an anonymous referee. We alone are responsible for the views expressed in this paper.

[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. For a lucid description of the card industry see Hunt (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 (for example 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.

ability of merchants to set different prices depending on the means of payment employed, such as credit cards, 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 card payments, even though merchants may face higher costs for card transactions due to EPN fees.[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 several policy questions. First, it helps assess 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). Second, when repeal of the NSR is not an option – because merchants' reluctance to surcharge derives from other characteristics of the trading environment – our analysis helps to understand the effects of pricing practices such as interchange fees and 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.

Some institutional and literature background will be helpful. Any payment network intermediates between consumers and merchants. In a *proprietary* (or "closed") network such as American Express, the same entity deals with, and sets the fees to, merchants and card users. In an *association* such as Visa (or "open network" since membership is open to multiple banks), a typical transaction involves two different banks: the cardholder's ("issuer") and the merchant's ("acquirer"). The issuer sets the fees to its cardholders (for example, a per transaction fee or, more often, a rebate) and the acquirer sets the fee to its merchants (the "merchant discount" – the transaction amount minus what the merchant receives from the acquirer). The association sets an *interchange fee* paid by the acquirer to the issuer, which of course affects their respective charges to merchants and cardholders. Formal economic analysis of the interchange fee was pioneered by Baxter (1983), who showed that the socially optimal fee must reflect the net benefits from card use on both sides of the transaction. However, in Baxter's analysis the association is indifferent between any levels of the interchange fee, because issuers and acquirers are assumed perfectly competitive.

Analyses of the interchange fee under imperfect competition among association members came considerably later. Schmalensee (2002) studies how the fee affects marketing efforts (and pricing) by issuers and acquirers. Closer to our focus are Rochet and Tirole (2002), who also provide a broader review of recent literature. In their model, consumers have unit demands for transactions but heterogeneous private values of paying with cards versus the outside instrument, "cash". Duopolist merchants are spatially

[3] Surcharges on credit card transactions were prohibited by federal statutes from 1968 to 1985 and remain prohibited by some states (for example, 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] 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.

differentiated and, given the interchange fee, choose simultaneously whether to accept cards and, then, set prices. Acquiring banks are assumed perfectly competitive, while issuers are imperfectly competitive.[5] Issuers' fee to cardholders is represented by a reduced form, decreasing function of the interchange fee, since a higher interchange fee lowers issuers' net marginal cost of issuing cards. If merchants can freely surcharge for card transactions, the level of the interchange fee is irrelevant ("neutral") and card diffusion among consumers is socially too low, because of the imperfect competition among issuers. With a no surcharge rule, there exists an equilibrium in which both merchants accept cards, the net price falls to card and rises to cash users, and card diffusion either remains too low (if so, the NSR raises overall welfare) or becomes excessive (if so, welfare may rise or fall).

Rochet and Tirole – and all other literature to our knowledge – focus on consumers' choice between cards and cash but hold the total quantity of transactions fixed.[6] Though a sensible first approximation, this is obviously an abstraction. Our analysis is complementary by allowing transaction quantities to vary. To focus on this dimension, we make two simplifications.

First, we treat payment mode as exogenous: one group of consumers (of mass one) use only cards, while others (of mass α) use only cash.[7] Transactions demand per capita is the same for members of each group and is downward sloping. Thus, while the number of card users is exogenous in our model, the per capita quantities of card and cash transactions are determined endogenously. The assumption that some consumers must use cash can be justified partly on grounds that a sizeable portion of the population are ineligible for cards.[8] Moreover, as shown by Rochet and Tirole (2002, Proposition 6), the key property that the NSR allows an EPN to tax cash users continues to hold when there is imperfect substitution between cash and cards. Our assumption that cardholders cannot use cash is strong. However, it introduces consumer benefits from cards in a simple (albeit extreme) way and, more importantly, it captures the opposite polar case to the one assumed in existing literature. There, a merchant who refuses cards only loses transactions to another merchant who accepts cards (Rochet and Tirole, 2002) or converts all customers to using cash (Wright, 2003). In practice, a merchant refusing cards can lose some transactions entirely, most obviously because some cardholders are liquidity constrained or perhaps would only buy certain items with cards due to other bundled properties of the card such as

[5] At least in the U.S., competition is viewed as stronger on the acquiring side. Rochet and Tirole (2002, p. 552) state that acquiring is "widely viewed as highly competitive," citing Evans and Schmalensee (1999), while issuing "is generally regarded as exhibiting market power."

[6] In Rochet and Tirole, this follows because of the assumptions of unit demands and that the market is always covered in the Hotelling competition between merchants.

[7] 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.

[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 superior dispute-resolution protection. Reality is likely to lie between the two polar cases.

Our second simplifying assumption is that merchants are local monopolists, so we do not explore an important aspect studied by Rochet and Tirole (2002) and Hayashi (2005), the interdependence between merchants' decisions to accept or reject cards. In Rochet and Tirole, imperfect merchant competition is the sole reason why under the NSR card use relative to cash use can be excessive (a merchant's gain from accepting cards comes partly by diverting sales from the other merchant). In our model, the NSR has a different welfare reducing aspect: it lowers per capita transactions by cash users due to elastic demands. Our environment leads to some interesting differences, for example, if rebates to card users are not feasible, then in our model the NSR can harm even *card* users.

Our results contrast sharply with those of Wright (2003), who finds that when merchants are local monopolists (as in our model), the NSR lowers price to card users and leaves price to cash users *unchanged*.[9] The absence of harm to cash users, and the strong prediction that the NSR unambiguously increases welfare, both hinge on his assumption of unit demand by cash users. More typically, when demand facing a merchant varies continuously as in our model and Rochet and Tirole's, the NSR induces the merchant (in part) to raise its price to cash users.[10] Holding other fees constant, this effect induces a cross-subsidy from cash to cards.

We follow Rochet and Tirole and others in assuming that acquiring banks are perfectly competitive, and we consider two polar cases regarding issuers. Section 6 analyzes almost perfect competition – homogeneous Bertrand issuers with prices set in discrete units. Our main model, however, assumes perfectly collusive issuers; coupled with perfectly competitive acquirers, the association will behave like a monopolist proprietary EPN that sets fees to cardholders and to the merchant. Since the merchant also is a monopolist, this environment yields double marginalization if the merchant can surcharge card transactions.

Given double marginalization, one might expect the NSR to increase consumer surplus and overall welfare. Like maximum resale price maintenance (RPM), that is known to be beneficial in such situations (Tirole, 1988), the NSR curbs the merchant's margin on card sales. The analogy, however, is imperfect because RPM only reduces the price of the targeted product, while the NSR squeezes the merchant's margin indirectly by requiring that the same price be charged for the other product (here, cash transactions), thereby causing that price to rise.[11]

---

[9] Wright also considers *homogeneous Bertrand* competition among merchants. As expected, the NSR cannot harm cash users then either, this time because merchants will specialize in accepting cash *or* cards.

[10] In Wright's model, consumers vary in their benefits from cards relative to cash but have identical maximal value v for the good if using cash. Moreover, there is a mass of consumers who will always use cash (as in our model) because their benefits from cards are less than the marginal cost of card services. A local monopolist merchant able to surcharge for cards will set a price v to cash users and a higher price to card users. Under no surcharging, and if v is sufficiently large, the merchant sets a uniform price of v because any higher price entails losing *all* cash transactions. The difference in our model is that cash users' demand for transactions is smooth, so under the NSR some price increase to cash users will be profitable. A smooth demand by cash users also arises in Rochet and Tirole, despite individuals' unit demands for transactions, because each merchant faces consumers that are spatially differentiated as well as smoothly distributed in their valuations for cards relative to cash.

[11] Indeed, because the NSR distorts the merchant's behavior in the other market, it cannot be used to completely eliminate the merchant's card margin (as the merchant would then drop cards), so the EPN would seemingly prefer RPM to the NSR. However, implementing RPM would be daunting for the EPN, as it would require dictating an enormous number of retail prices.

Despite the rise in the cash price, one might still expect the NSR to increase overall welfare as with optimal taxation (or Ramsey pricing) where inefficiency is reduced by using a broader tax base to lower the tax rate. The (imperfect) analogy is that, holding constant the EPN's fees, an NSR would lead the merchant to set an intermediate *uniform* price for all transactions instead of a higher card price and lower cash price, thus reducing misallocation in the mix of transactions. Again, however, the analogy is flawed. Since the EPN is unregulated, it will alter its fees when allowed to tax non-card sales via the NSR. If rebates from issuing banks to card users are not feasible, the EPN raises its fee to the merchant so much that the price paid by *card* users also can rise.[12] If rebates to card users are feasible, the NSR can induce excessive use of cards relative to cash – reversing the no NSR bias. We show how the welfare tradeoffs depend on two parameters: the size of the cash users' group relative to card users, our parameter α, and the merchant's additional benefit from handling a card rather than a cash transaction, captured by a parameter *b*.

Finally, the analysis brings out sharply how the NSR, by constraining merchant pricing, breaks the EPN's indifference between charging to merchants or to card users. As several authors have noted (and we show as well), if merchants can surcharge card transactions, the division of the EPN's charges is neutral – only the total charge matters. With the NSR, however, the EPN prefers to load its charges on merchants, in the case of a card association by charging a high interchange fee. Indeed, the EPN then prefers *negative* fees to card users – rebates. Such rebates are often viewed as reflecting the inability of an association to prevent its issuing banks from competing for card users and dissipating rents generated by high fees to merchants. Our analysis reveals a different possibility: rebates allow an EPN to better exploit a no-surcharging constraint on merchant pricing. In our model, the EPN grants rebates to card users so as to boost their demand and raises its fee to merchants knowing that they will absorb part of the increase, because under the NSR any price increase must apply equally to cash users.

The paper is organized as follows. Section 2 presents the model and shows that the division of the EPN's fees between card users and merchants (hence also the interchange fee) is neutral when the merchant can set different cash and card prices. Section 3 shows that with the NSR, the EPN prefers to shift its charge away from card users. Section 4 addresses the case where rebates to card users are not feasible. If the cash group is relatively small, the NSR harms even card users. With a larger cash group, card users gain but aggregate consumer surplus is still lower than under no NSR. Total surplus, however, is higher if and only if the cash group is sufficiently large.

Section 5 allows for rebates. With the NSR, the EPN always grants rebates. Relative to no NSR, card users gain but cash users lose; aggregate consumer surplus (of cash plus card users) rises only if the cash group is sufficiently *small* whereas total surplus rises only if the cash market is sufficiently *large*. A larger merchant benefit from using cards instead of cash improves the effects of the NSR on both total surplus and overall consumer surplus when card user rebates are feasible – but the reverse occurs when rebates are not feasible.

Section 6 considers the case of Bertrand rather than collusive issuers. An NSR again will induce rebates, benefiting card users but harming cash users. Overall consumer surplus rises regardless of the relative sizes of the two groups. However, provided the

[12] In contrast, standard comparisons of uniform pricing vs. third-degree price discrimination by a monopolist (our merchant) find, under regularity conditions that are met here, that requiring a uniform price causes at least some price(s) to fall (Nahata et al., 1990; Malueg, 1992).

merchant's benefit from card use is not too large, overall welfare falls with the NSR. Section 7 concludes.

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

$$U(p_c, q_c)=V(q_c)-p_c\, q_c,$$

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

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 for the merchant's good is given by $V'(q_j) = p_j$.

Merchants: Merchants are local monopolists. 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 merchant is charged a per-unit fee i by the EPN.

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 quantity 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 the *total tax* imposed on the card market by the EPN and the term $i+t-b$ as the merchant's *net marginal cost* in the card market.[13] Thus, the merchant's optimal quantity in the card market is only a function of the total $i+t-b$ and not of the composition of the charges. For given $(i,t),$ we denote the value of the merchant's 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 $(i,t),$

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

[13] Positive b could be reinterpreted as an upward shift in card users' inverse demand relative to cash users. To see this, rewrite the profit function as $(V'(x)+b-(i+t))x$.

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

Electronic Payment Network: 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. A card association will behave in the same way under two 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 acquirers will earn zero profit and will fully pass through any variation in the interchange fee. We therefore suppress the distinction between the interchange fee and merchant discount, and view the issuing banks as setting the fee to the merchant *i*, through their interchange fee, to maximize issuers' profits. Condition (b) implies that charges to card users also are chosen to maximize overall profits of issuing banks.[14] 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.[15] 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 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 neutrality result.[16]

Proposition 1: Suppose merchant surcharging for card transactions is allowed. Then equilibrium card transactions, merchant profit and EPN profit all depend only on the EPN's total fee, $i+t$, and not on i and t individually. That is, if (i,t) maximizes the profits of the EPN, then so too does any pair (i',t') where $t'+i'=t+i$.

[14] However, in Section 6 we analyze the other polar case of competition among issuer banks – card user fees are then set in a Bertrand fashion.

[15] 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.

[16] 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).

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 proofs and intuition for our results are most concisely conveyed for the case of linear demand. Thus, the remainder of the paper restricts attention to this case:

A1) *Consumer per capita inverse demand is $V'(x)=1-x$ and the merchant's benefit from card use satisfies $b<1$.*

The assumption of linear demand enables closed form solutions for most of the relevant variables in the analysis, and ensures that properties **P1)-P3)** hold – many of our qualitative results hold for any demand curves that satisfy these properties:

P1) The merchant's revenue is strictly concave in quantity and price and any increase in the merchant's marginal cost in serving cards is not fully passed through to consumers when surcharging is possible.

P2) The EPN's revenue function, $x(x\,V''(x)+V'(x)+b)$ is strictly concave in quantity.

P3) With merchant surcharging of card transactions, the EPN sets a total fee, $i+t>b$.

P3) is a mild restriction. With linear demand, the EPN's optimal total charge is $i+t=(1+b)/2$. For this to exceed b requires $b<1$, the merchant's added benefit from card use (for example saving on cash handling cost) is less than the consumers' maximal willingess to pay for the good itself. Under **P3),** the merchant's net marginal cost is positive for serving card users, but zero for cash users. Standard revealed preference arguments (for example Tirole, 1988, pp. 66-67) imply that as marginal cost rises, the merchant's monopoly quantity falls, hence per capita transactions will be lower for card users than for cash users when the merchant can surcharge for card transactions.

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 for card than for cash transactions, $p_e^M \le p_c$.[17] By **P3)**, when merchant surcharging is allowed, the EPN's optimal aggregate fee causes the merchant to face a higher net marginal cost of serving card than cash users; thus, when the two demand curves facing the merchant are equal – as occurs when $t = 0$ – the merchant prefers a higher price to card users ($p_e^M > p_c$). The NSR would then bind on the merchant. Proposition 2i) below shows that the NSR will also bind if, instead, the EPN's fee to card users is positive but below some threshold.[18]

[17] 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.

[18] An implication of this observation is that, with a low card user fee (which Proposition 2 shows is desired by the EPN), we can formulate the no-surcharge rule mathematically as the inequality constraint, $p_e^M \le p_c$ 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

Moreover, it is costless for the EPN to adopt such a profile of charges pre-NSR since, by Proposition 1, only the sum of charges then matters. Proposition 2ii) shows that if the binding NSR is accepted, the EPN would benefit. Proposition 2iii) shows that under an NSR the EPN has the incentive to continue raising the charge to the merchant and lowering it to card users.

Proposition 2: *Fix $i+t$ at the level $k>b$ 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;

(ii) If an NSR is accepted, holding (i,t) fixed, then cash purchases fall but card purchases rise and, hence, EPN profits also rise;

(iii) Provided the merchant continues to accept the NSR, a cut in the card user fee t and an equal rise in merchant fee i increases per capita card transactions and EPN profits.

The intuition for Proposition 2ii) is straightforward. With a binding NSR, the merchant will choose a uniform price between its pre-NSR card and cash prices: starting from a uniform price equal to the optimal card price 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. The EPN therefore gains if a binding NSR is accepted: EPN profits rise at the pre-NSR charges (i,t) and any departure from these price post-NSR, by revealed preference, would further benefit the EPN.

The intuition behind Proposition 2iii) 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 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). With surcharging, this indeed would be the outcome, hence the familiar neutrality result. But since the NSR forces the merchant to charge the same price to cash users as to card users, the merchant prefers to raise her uniform price by less than the full increase in i and accept a lower margin on card sales.[19]

---

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. Proposition 2i) shows when the effects of the two constraints are the same.

[19] This argument also 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 change 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 change 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 under the NSR ($\partial p/\partial t$ is negative but smaller in absolute value, so $1+\partial p/\partial t$ is positive and larger): card users receive this cut directly, and (by **P1)**) the merchant responds by increasing p by less than with no NSR.

Proposition 2 shows that with the NSR, it becomes relevant how $i+t$ is distributed: the EPN prefers a lower card user charge (provided the merchant still accepts). Rebates to card users – negative t – are often taken as evidence of the inability of a bank card association to control competition for card users by its member banks. Proposition 2iii) offers an alternative interpretation: rebates can 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 determines the floor on t? One limit may be institutional. For historical, practical or regulatory reasons, rebates to card users may not be feasible.[21] Section 4 investigates this case. A priori, the binding constraint then may be the non-negativity of t, the merchant's IR constraint, or both.[22] Proposition 3 shows, however, that the non-negativity constraint always binds. Section 5 allows rebates, showing that the EPN may be constrained either by the merchant's IR constraint or by the need to ensure that the merchant will not price so high that cash users are driven out (a type of incentive compatibility 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 $(i_0,0)$, the IR does not bind, then, by Proposition 2, these prices are optimal for the EPN. If the IR is violated at these prices, in Proposition 3 we provide sufficient conditions under which setting $t = 0$ is still optimal for the EPN.

With an NSR, cash users and card users pay the same merchant price, p. Linear demand then implies that the per capita transactions are $q_c=1-p$ for cash users and $q_e=1-p-t$ for card users. For any given (i,t), the merchant then selects price to solve

$$\Pi^{NSR}(i\ t;b) \equiv max_p\ \alpha p(1-p)+(p-i+b)(1-p-t),$$

yielding

$$p(i,t;b)=\frac{1+\alpha+i-b-t}{2(1+\alpha)},\quad q_c(i,t;b)=\frac{1+\alpha-i+b+t}{2(1+\alpha)},\quad q_e(i,t;b)=\frac{1+\alpha-i+b-t(1+2\alpha)}{2(1+\alpha)} \quad (3)$$

[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). By the late 1990s, roughly half of all credit volume were 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.

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

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

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

$$t \geq 0 \qquad \textbf{(No Rebates)}.$$

The objective function of the EPN is concave in *(i,t)* while the IR constraint is linear in *(i,t)*. Thus the Kuhn-Tucker first order conditions are sufficient.[23]

Proposition 3 (Prices): *Suppose card rebates are not feasible ($t \geq 0$). Under the NSR:*

(i) For any relative size of the cash market, α, the EPN chooses 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 α > α*.*

(iii) If α > α, then i falls as α rises and i+t-b (=i-b) is independent of the merchant benefit b.*

(iv) If α < α, i increases in α.*

(v) For all α, i is higher than the total charge under surcharging

Proposition 3i) shows that the EPN's desire for lower card user fees and higher merchant fees illustrated in Proposition 2 drives user fees to zero even if the merchant's IR constraint binds before the EPN achieves its optimal fee pair. 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 3ii) shows that the merchant IR constraint binds if and only if the cash market is not too small. (With *b=0*, the IR binds if $\alpha > \alpha^* = 1/3$.) Since, the merchant's profit from serving only cash customers is increasing in the size of the cash market, as the latter increases in the range where the merchant's IR binds, the EPN must reduce *i* to maintain merchant participation. By contrast, as *α* increases from *0* to *α**, the EPN responds by *raising i*, because in this range the IR is not binding and a larger cash market reduces the merchant's pass-through from *i* to the uniform retail price. Finally, observe that the NSR affects not only the EPN's fee *structure* but also the *level*: the EPN's total fee is higher under the NSR for all *α* (Proposition 3iv)).

For Proposition 4ii)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.

[23] With the exception of Propositions 3iv), 3v) and 4ii)c), Propositions 3 and 4 can be shown to hold for more general demand functions than linear, that is, those which satisfy **P1)-P3**.

Proposition 4 (Quantities and Welfare): *Suppose an NSR is imposed but card user rebates are not feasible ($t \geq 0$). Let CS denote Consumer Surplus. Compared to the equilibrium with no NSR,*

(i) If the cash market is small enough that the merchant IR does not bind ($\alpha < \alpha^*$), then:

a. Cash users' transactions and CS are lower;

b. Card users' transactions and CS are unchanged if $b = 0$ and lower if $b > 0$;

(ii) If the merchant IR binds ($\alpha > \alpha^*$), then:

a. Cash users' transactions and CS are lower;

b. Card users' transactions and CS are higher if α is sufficiently larger than α^*;

c. Aggregate quantity ($q_e+\alpha q_c$) and aggregate CS are lower for all α and all b.

d. ΔTS^{NR} rises in α and falls in b. For $b=0$, $\Delta TS^{NR}=0$ at a value of α above α^*.

We first explain the intuition underlying Proposition 4i) and then 4ii).

IR Not Binding ($\alpha < \alpha^*$). 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. This indeed would occur if the EPN's fees remain fixed.[24] But 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 $b>0$ (Proposition 4i)).

Therefore, when the merchant's IR is not binding, the welfare consequences of the NSR are stark. The NSR then reduces even card transactions (leaving them unchanged only if $b=0$), thus harming card users. 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.

IR Binding ($\alpha > \alpha^*$). In this case, the NSR still reduces the merchant's profit – since the merchant 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 4ii)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 $-(1+\alpha)Q^2V''(Q) = \alpha x_0 V'(x_0)$ or

[24] Under cost and demand conditions satisfied here, 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).

$$- Q^2V''(Q) = (\alpha/(1+\alpha))x_0V'(x_0). \quad (4)$$

As the size of the cash market increases to satisfy (4), the EPN must induce an increase in Q (concavity of the merchant's revenue function in quantity implies merchant profit is increasing in Q), 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 **P3)**).

Total quantity is lower under the NSR (Proposition 4ii)c)). 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 constant, but in fact the EPN raises its total charge (Proposition 3iv) to exploit the decreased elasticity of demand it faces from the merchant, so total quantity falls. Overall consumer surplus, therefore, also must fall because of the following property:

Lemma 1: *Consider any pair of prices (p_c, p_e) to cash users and card users (where p_e includes any EPN charge t) that yield a fixed total quantity of transactions, $\alpha q_c + q_e = k$. Then overall consumer surplus of cash and card users increases with the dispersion in per capita quantities, $|q_c - q_e|$.*

Lemma 1 can be understood as follows. Identical linear demands for cash and card users imply that (a) total quantity is constant only if the weighted average price $\alpha p_c + p_e$ is constant, and (b) dispersion in per capita quantities $|q_c - q_e|$ is linear in $|p_c - p_e|$. Overall consumer surplus, $\alpha S(p_c) + S(p_e) = (1+\alpha)[S(p_c)\alpha/(1+\alpha) + S(p_e)/(1+\alpha)]$, is proportional to the consumer surplus of an individual who faces p_c and p_e with probabilities $\alpha/(1+\alpha)$ and $1/(1+\alpha)$. An individual's consumer surplus $S(p)$ is convex in price, so any mean-preserving spread of the prices will increase the expected surplus. Since the NSR with no rebates *reduces* the spread in per capita quantities (to zero) as well as total quantity, overall consumer surplus must fall (Proposition 4ii)c)).

Total surplus, however, can be higher with the NSR if the cash market is large enough. 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 gain from improved allocation outweighs the loss from the fall in total quantity so the NSR raises total surplus.

There are two reasons why ΔTS^{NR} increases in α for $\alpha > \alpha^*$ (Proposition 4ii)d). First, the merchant's preferred price is lower for cash than for card users. Thus, as the cash market becomes relatively more important (as α increases), for given EPN fees the merchant's optimal *uniform* price falls. Second, to respect the merchant's IR, the EPN must cut its total fee as the cash market grows. For both reasons, 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. (The reason why ΔTS^{NR} falls in b, as stated in Proposition 4ii)d), is discussed in Section 5.)

5 Equilibrium when rebates are feasible

Proposition 3i) 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.[25] 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 arise with $t \geq 0$ – (per capita) inverse demand of card users is then lower than that of cash users, so any price that yields cash sales also yields card sales – but must be tackled under rebates.

Propositions 5 and 6 compare the equilibrium under the NSR with rebates to that under no NSR. Proposition 7 summarizes the incremental effect of rebates by comparing the equilibria under the NSR with and without rebates.

Proposition 5 (Prices): 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 5i) is illustrated in Figure 1 for 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 Proposition 2 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 $i+t = i_0$ (that is, a cut in t and an equal increase in i) raises EPN profit. Expression (4) 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.

[25] A monopolist that faces two markets but is prohibited from 3rd-degree price discrimination will drop the low market if the dispersion in the demands is sufficiently large (Tirole 1988, p.139).

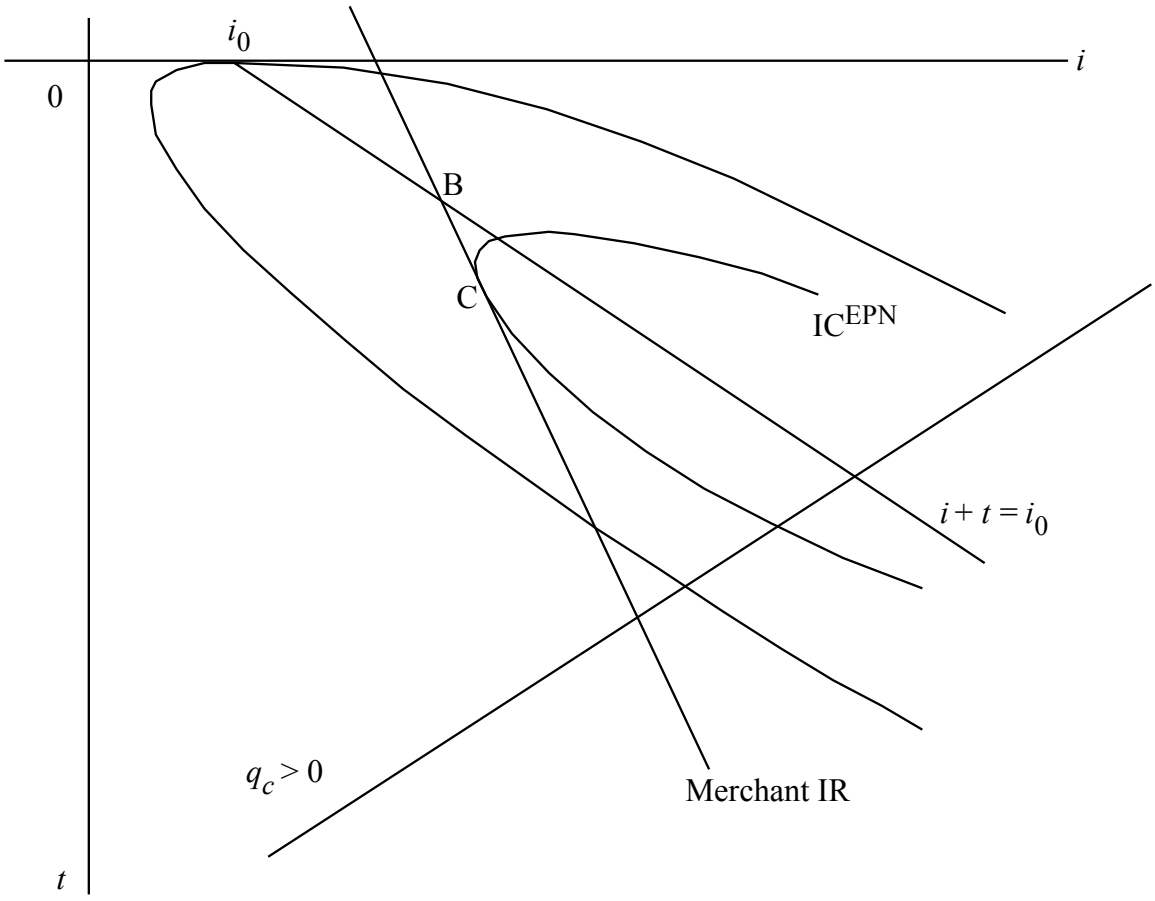

Figure 1

Proposition 5ii) shows that even with rebates feasible, the EPN cannot always fully extract the merchant's surplus. If the cash market is sufficiently small, the floor on t is not the merchant's IR constraint but the need to induce the merchant to continue serving cash users.

When the cash market is large enough that EPN charges are determined by the IR constraint, Proposition 5iii) shows that 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; however, the spread between i and t (which is irrelevant under surcharging) shrinks as α increases (Proposition 5iii). The EPN's equilibrium fees when the IR binds are (see Appendix),

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

The total charge, t^*+i^*, is therefore $(1+b)/2$, the same as under surcharging, but lower than under the NSR with no rebates. The EPN prefers to grant rebates and accept a reduction in $i+t$, as needed to satisfy the merchant's IR, because rebates are a more effective way to boost card transactions. (Recall from Proposition 2, under the NSR a reduction in t and equal increase in i would increase card transactions. See also fn. 19) As the cash market grows, the EPN meets the IR with the same total charge (instead of cutting it as under no

rebates) by reducing the spread between i and t: t^* rises with α (smaller rebates) while i^* falls. The merchant benefits from this reduced spread: 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 6 (Quantities and Welfare): *Suppose an NSR is imposed and rebates are feasible. For α large enough 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 6 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 6iv). 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 per capita transactions of cash versus card users. The efficient per capita levels are 1 for cash and $1+b$ for cards. With surcharging, the equilibrium 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 $1/2$.[26] As α increases, therefore, the allocation improves only under the NSR, so ΔTS^R rises in α. With zero merchant benefit from card use, $b=0$, ΔTS^R is positive if and only if $\alpha > 1/3$. The case of $b>0$ is discussed shortly.

By contrast, the change in aggregate consumer surplus when moving to the NSR with rebates is less favorable as α increases (Proposition 6v). The narrowing of the gap between per-capita cash and card quantities as α increases occurs only under the NSR, and is

[26] The card quantity rises because the merchant's *preferred* price is lower to cash than to card users, and an increase in α magnifies the relative importance of cash users, leading the merchant under the NSR to cut its uniform price. This effect dominates an opposing effect: that as α increases, the EPN reduces the spread (i-t) in its fees (Proposition 5iv)), which of itself would reduce card transactions (by the reverse argument of Proposition 2iii)).

harmful to overall consumer surplus, by Lemma 1. For $b=0$, the NSR with rebates reduces overall consumer surplus if and only if $\alpha > 1/3$.

Positive Merchant Benefit from Cards. Moving from surcharging to the NSR with rebates, therefore, causes opposite changes in total surplus and overall consumer surplus if $b=0$: when $\Delta TS^R > 0$ ($\alpha > 1/3$), $\Delta CS^R < 0$. The increase in EPN profit then necessarily comes at least partly by harming other parties (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 6vi) and since both equal 0 at $\alpha = 1/3$, for $b>0$ there is an interval around $\alpha=1/3$ in which ΔTS^R and $\Delta CS^R > 0$. These results are illustrated in Figure 2.

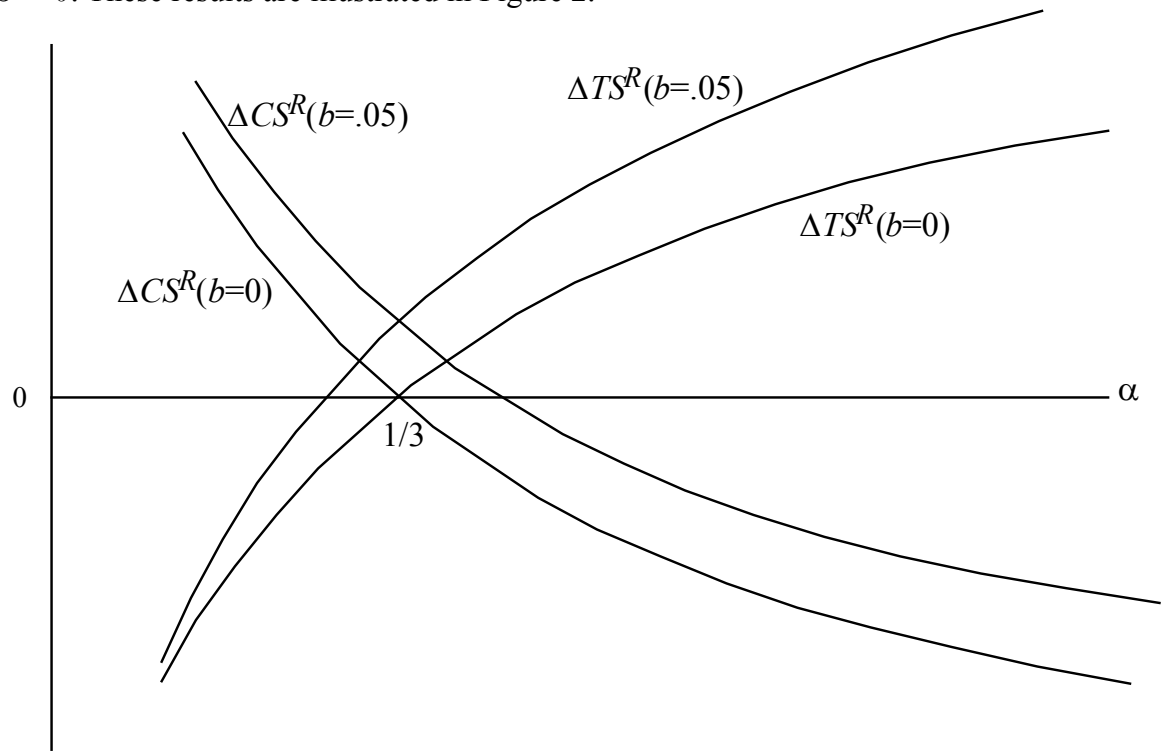

Figure 2

The intuition for why ΔTS^R and ΔCS^R increase 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 card and cash quantities are $q_e^{**} = 1+b$, $q_c^{**} = 1$, hence $q_e^{**} - q_c^{*} = b$. With surcharging, quantities are $q_e = (1+b)/4$, $q_c = 1/2$, hence $q_e - q_c = (b-1)/4$. Recalling that $b<1$, under surcharging $q_e < q_c$, and the gap closes at the rate $\Delta b/4$. With the NSR and rebates, per capita transactions are higher for card than for cash users. Moreover, the difference between the card and cash quantities increases faster with b under the NSR than under surcharging and never exceeds the efficient gap, b. Thus, an increase in b magnifies the allocation advantage of the NSR. Finally, ΔCS^R increases in b. The total quantity of transactions is the same under surcharging and the NSR with rebates. But the spread in per capita quantities, $| q_e - q_c |$, rises with b under the NSR with rebates but falls under surcharging, and Lemma 1 shows that overall consumer surplus increases with the spread.

Interestingly, the NSR's effect on total surplus is *less favorable* the higher is b in the case when rebates are not feasible and α in the range where the merchant's IR binds (Proposition 4ii)d, ΔTS^{NR} falls in b). Since the distortion from double marginalization on card transactions increases with b under surcharging (the gap between the efficient and actual card quantities is $3(1+b)/4$), and since curbing double marginalization is what motivates the NSR in our model, one might expect the NSR to be better for total surplus the larger is b. With no rebates, however, the EPN under the NSR responds to an increase in b by raising its charge to the merchant enough that the net marginal cost of card transactions, $i+t-b$, stays constant (Proposition 3iii), and therefore quantities are unchanged.[27] Under surcharging, the EPN responds to an increase in b by reducing $i+t-b$, and therefore card transactions rise. 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 allows card transactions to rise *faster* as b *increases* than under no NSR, so ΔTS^{R} rises in b.

Proposition 7 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 enough that the merchant's IR determines EPN pricing with or without rebates.

Proposition 7 (NSR, Rebates vs. No Rebates): *Suppose the NSR is imposed, the merchant's IR binds, and rebates are feasible. Compared to the outcome under the NSR with no rebates:*

Card users' consumer surplus is higher, cash users' consumer surplus is lower, and aggregate consumer surplus is higher with rebates.

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 7ii) reflects the ability of rebates under the NSR (and only then) to more effectively reduce double marginalization than by only cutting the fee to the merchant. Recall that with the NSR an increase in i by 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 (3), $\Delta p = 2\Delta/(2(1+\alpha)) = \Delta/((1+\alpha)$.) Moreover, when cutting t below 0, the EPN raises i by less than the rebate amount to respect the merchant's IR constraint. Since the aggregate EPN charge $i+t$ is lower with rebates than without, total transactions are higher. Overall consumer surplus therefore also is higher with rebates: total quantity is higher and so is the spread in per capita quantities (under the NSR, the spread is positive with rebates but zero without). Finally, total surplus is 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).

Cash users, however, lose from the NSR even with no rebates to card users (Proposition 4), and lose further if rebates are feasible. Granting rebates increases the

[27] Under the NSR and no rebates, when the merchant'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.

inverse demand of card users, prompting the merchant to raise its retail price.[28] In addition, when the EPN grants rebates, it also raises its fee to the merchant somewhat, putting further upward pressure on the merchant's price (see (3) where p increases in i and decreases in t, while q_c does the reverse).

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 card users, 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 card users 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, $\varepsilon \approx 0$. By the standard Bertrand logic, the equilibrium t_W satisfies $t_W = -i + \varepsilon$. 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 (3) 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 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 ($\varepsilon>0$), the EPN will wish to generate the largest possible quantity of such transactions. Since card

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

[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 $t<-b/2$.

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 8 summarizes the effects of this incentive on equilibrium quantities under the NSR and competitive issuers.[30]

Proposition 8: *Assume b=0. With perfectly competitive issuers, in the equilibrium under the NSR:*

(i) If α<1, the EPN sets i until merchants are just indifferent between selling to cash customers or not; if α >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 8i) 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 8ii)) because demand is linear and the total EPN fee remains the same (ε). 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 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 (Lemma 1).

Result 8iv) 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 level. 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.

[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.)

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 card and cash prices independently). Throughout, the NSR benefits the EPN but harms cash users and the merchant. Other welfare effects depend on the ratio of cash to card users, the merchant's benefit from card versus cash transactions, and whether rebates to card users are feasible.

If rebates are not feasible, the EPN charges zero to card users but sets its merchant fee above the total fee that it charges with no NSR. This increase in total fee under the NSR reduces total transactions and aggregate consumer surplus; with a sufficiently small cash market, even card users pay more. 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.

The above results apply for the case of monopoly pricing at both the merchant and EPN levels. However, 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.

In order to focus on how the NSR affects transaction levels per consumer, our analysis abstracted away from consumers' choice of the means of payment. Extensions of this research would include analyzing the effects of the NSR when both transaction volumes and consumer choice of the means of payment are endogenous, and under a broader class of merchant market structures (for example, 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 (for example, Rochet and Tirole, 2003).

8 References

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

Carleton, Dennis and Alan Frankel (1995a) "The Antitrust Economics of Credit Card Networks," *Antitrust Law Journal*, 63: 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: 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*, 2: 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: 861-901.

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

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, Joshua. and Stephen King. (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: 872- 886.

Hayashi, Fumiko (2005) "A Puzzle of Card Payment Pricing: Why are Merchants Still Accepting Card Payments?" Payments System Research WP04-02, December 2004, Federal Reserve Bank of Kansas City, September.

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

Malueg, David (1992) "Direction of Price Changes in Third-degree Price Discrimination: Comment," Working Paper 92-ECAN 95, Freeman School of Business, Tulane University.

Nahata, Babu, Krzysztof Ostaszewsi, and Prasanna Sahoo (1990) "Direction of Price Changes in Third-degree Price Discrimination," *American Economic Review*, 80: 1254-1258.

The Nilson Report, May, 2003.

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

Rochet, Jean-Charles. (2003) "The Theory of Interchange Fees A Synthesis of Recent Contributions," *Review of Network Economics*, 2: 97-124.

Rochet, Jean-Charles and Jean Tirole (2002) "Cooperation Among Competitors: Some Economics of Payment Card Associations," *Rand Journal of Economics*, 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*, 1: 85-96.

Schmalensee, Richard (2002) "Payment Systems and Interchange Fees," *Journal of Industrial Economics*, 50: 103-122.

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

Wright, Julian (2003) "Optimal Card Payment Systems," *European Economic Review*, 47: 587-612.

9 Appendix

Proof of Proposition 2: (i) Since merchant sales are decreasing in marginal cost, $x(k-b) < x_0$ 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 $V(\cdot)$. 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) Let $f(p)$ be 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 3: (i) When the IR constraint does not bind, the result follows from Proposition 2iii). Now suppose the IR binds and consider (t,i) space. At $t = 0$, and i such that the merchant IR curve binds, we show that the slope of the EPN level set is a lower negative number than the slope of the merchant IR curve which has slope with absolute value less than one. This implies that this point is a constrained maximum.

Under an NSR, the Lagrangian representing the merchant's profit maximization problem is

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

where $\lambda>0$ is the lagrangian on the No Surcharge constraint. The EPN's profit function is given by

$$\Pi^e=(i+t)q^e(t,i;b).$$

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

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

where $Q \equiv q_c = q_e$ (at $t = 0$ with an NSR). Note that the denominator in the first expression is non-negative since the EPN's profit function is quasi-concave along the line $t=0$ and the fact that the IR constraint binds implies it is constrained to select an i less than its unconstrained optimum. Equation (3) which provides the merchant's optimal choice of q^e then implies

$$\frac{\partial q_e}{\partial t} = -\frac{1+2\alpha}{2(1+\alpha)} \; < \; \frac{\partial q_e}{\partial i} = -\frac{1}{2(1+\alpha)} \; < \; 0.$$

The first order conditions for the merchant's optimal choice of quantity at $t = 0$ imply that $i = b + \lambda\,(1+\alpha)/\alpha$. This yields

$$i\left(\frac{\partial q_e}{\partial i} - \frac{\partial q_e}{\partial t} \right) \; = \; b\left(\frac{\alpha}{1+\alpha} \right) + \lambda \; \geq \lambda$$

and

$$0 \leq i\frac{\partial q_e}{\partial i} + Q \; = \; b\frac{\partial q_e}{\partial i} - \frac{\lambda}{2\alpha} + Q \; \leq Q - \frac{\lambda}{2\alpha}.$$

The first inequality comes because the IR is binding on the EPN and the second inequality follows because q_e is decreasing in i. Combining these results yields

$$\frac{i\left(\frac{\partial q_e}{\partial i} - \frac{\partial q_e}{\partial t} \right)}{i\frac{\partial q_e}{\partial i} + Q} \geq \frac{\lambda}{Q - \frac{\lambda}{2\alpha}}.$$

Now consider the level sets of the EPN and the merchant in *(t,i)* space. Subtracting the second from the first yields, after substituting the inequality from above,

$$\frac{di^{e}}{dt}_{t=0} - \frac{di^{M}}{dt}_{t=0} \geq \left(\frac{\lambda}{Q - \frac{\lambda}{2\alpha}} \right) - \left(\frac{\lambda}{Q} \right) \; \geq \; 0.$$

Recalling that $Q - \lambda/2\alpha$ is positive, 3i) follows.

(ii) Note that at $t = 0$, the first order conditions for the merchant's problem 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 (4) in the text)

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

The right side is increasing in α. Concavity of the merchant revenue function in quantity 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) The merchant choice of Q is strictly decreasing in i-b, so, holding b fixed, 3ii) implies i decreasing in α. When the IR binds, Q is determined by (1A) which, in turn, determines i-b.

(iv) If the IR does not bind, 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 4: **(i)** Proposition 3 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. Concavity of the EPN's profit function 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 (1A) 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. **P3)** implies that eventually cardholder purchases and surplus are higher with the NSR.

c)-d) With an NSR and linear demand, merchant profit is $(1+\alpha)Q^2$ so (4) can be written $Q^2= \alpha/(4(1+\alpha))$. 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

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

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

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,

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

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

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

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 $b=0$.).

The change in consumer surplus is

$$\Delta CS = .5(Q_c+q_c)\Delta Q_T + .5((Q_e-Q_c)+(q_e-q_c))\Delta Q_e \ , \qquad (3A)$$

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 5: i) If $\alpha<\alpha^*$, then the IR does not bind under no rebates and, by Proposition 2iii), the EPN increases profits by holding i fixed and lowering t. If $\alpha \geq \alpha^*$, from the Proof of Proposition 3i) 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. Repeating (3), if the merchant prices to serve both markets, the solution is

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

The merchant chooses to serve both markets and thus selects prices and quantities as in (4A) if and only if

$$i-t-b\leq\sqrt{1+\alpha} \tag{5A}$$

Using the values for p, q_c,q_e from (4A) in (1A) gives the values of (i,t) for the EPN when the IR binds. Maximizing with respect to (i,t) yields

$$t^*=-\frac{1+b}{4(\alpha+\sqrt{\alpha}\sqrt{1+\alpha})}-\frac{b}{2},\quad i^*=\frac{1+b}{2}-t^* \tag{6A}$$

Equation (6A) implies that (5A) is violated as α gets small.
(iii) Follows from Equation (6A).

Proof of Proposition 6: (i)-(iii) Follow from (4A) 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 (2A) 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 (3A) 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 $-\alpha$, 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 7: (i) Equation (4A) shows that q_c falls as $i-t$ rises and q_e rises if $i+t$ and t fall. Propositions 3 and 5 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 4 and 6 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 (3A) 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 8: (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 (4A) 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 $1/2$.