THE JOURNAL OF INDUSTRIAL ECONOMICS 0022-1821
Volume LII March 2004 No. 1

THE DETERMINANTS OF OPTIMAL INTERCHANGE FEES IN PAYMENT SYSTEMS*

JULIAN WRIGHT†

This paper presents a model of a card payment system as a two-sided market that allows for partial participation by heterogeneous consumers and merchants. Taking into account the strategic effects arising from competition between merchants, the model is used to characterize the optimal structure of fees between those charged to cardholders and those charged to merchants and, more specifically, the level of the interchange fee that banks charge each other. The results modify the existing characterizations of the interchange fee, and explain the source of potential deviations between the privately and socially optimal level of the fee.

I. INTRODUCTION

TWO OF THE MOST IMPORTANT ADVANCES in the history of payment systems have been the development of debit and credit card payment systems over the last half-century. Such cards are typically offered through card associations, like MasterCard and Visa, that involve competing member banks issuing cards to consumers (these are known as issuers) and providing transaction processing to merchants (these banks are known as acquirers).[1] These are large systems. Visa involves 800 million cardholders, 27 million merchants, 21 thousand issuers and acquirers, and was used for U.S$2.3 trillion dollars of transactions in 2001 (Lyons, 2002). Despite the popularity

* Thanks go to Francis Bloch, Dennis Carlton, Henry Ergas, Joshua Gans, Stephen King, John Panzar, Michael Whinston, as well as other participants at seminars at the Bank of Finland, Federal Reserve Bank of Chicago, Federal Reserve Bank of Philadelphia, National University of Singapore, Northwestern University, U.C. Berkeley, University of Auckland, University of Canterbury, and the University of Toulouse for helpful comments. I would also like to thank two anonymous referees for their excellent comments, the Centre for Research in Network Economics and Communications (CRNEC) at the University of Auckland for funding research assistance, and Visa International for funding earlier related research (Wright, 2000). The views in this paper are solely my own, as are any errors.

† Author's affiliation: Department of Economics, Faculty of Arts and Social Sciences, National University of Singapore, AS2 Level 6, 1 Arts Link, Singapore, 117570.
email: jwright@nus.edu.sg

[1] These card associations contrast with proprietary card systems such as those offered by American Express and Discover, that charge cardholders and merchants directly.

of these systems, or perhaps due to it, the rules which govern them have recently been called into question.[2]

When consumers use debit or credit cards for a purchase, an interchange fee is paid from the acquirer to the issuer. The level of this interchange fee affects the relative fees faced by cardholders and merchants. A higher interchange fee raises the costs of acquirers, who will charge merchants more, and lowers the effective costs of issuers, who will charge cardholders less (or in fact, provide them with rebates). Frankel [1988] and, more recently, policymakers such as the European Commission and the Australian central bank, have argued that MasterCard and Visa have set interchange fees too high, the result being that merchants pay too much for accepting these cards, a cost which is ultimately passed on to their customers who pay with cash. Consumers who pay by credit cards, it is argued, are subsidized by taxing cash-paying customers, resulting in excessive card usage.[3]

This paper addresses what determines the optimal interchange fee (and so fee structure) in a card association, both from a private and a social perspective. This is interesting not only in order to analyze the policymakers' concerns about credit card systems, but also since the structure of fees in these card systems resembles that in other two-sided markets (firms advertising in Yellow Pages bear the costs of Yellow Pages rather than readers retailers located in shopping malls bear the costs of running the shopping mall rather than shoppers and employers listing jobs on a website bear the costs of the website rather than job-seekers).

Baxter [1983] provides the first formal analysis of interchange fees in a payment scheme. His analysis relies on three underlying assumptions: (1) issuers and acquirers are perfectly competitive and make no profit, (2) merchants do not accept cards for any strategic purpose (in particular, they do not accept cards to attract customers from rival merchants who do not accept cards), and (3) in working out the interchange fee implied by his analysis, implicitly it is assumed there is no variation in the benefits that merchants get from accepting cards. Assumption (1) implies the card schemes are indifferent to the level of interchange fees, and so cannot be used for the basis of a positive analysis of interchange fees. Assumption (2) ignores the business stealing motive for accepting cards, which can have an important bearing on both the privately and socially optimal interchange fee. Assumption (3) leaves unanswered how interchange fees should be set given heterogeneity across merchants.

[2] For useful background on the economics of payment cards and the associated policy and legal debates, see Evans and Schamlensee [1999], Chang and Evans [2000], and Chakravorti and Shah [2003].

[3] Despite this, proprietary schemes such as American Express and Discover, which also cover most of their costs by charging merchants rather than cardholders, have not been the subject of the same regulatory interest.

Schmalensee [2002] relaxes assumptions (1) and (3), emphasizing the need to balance cardholder and merchant demand by setting an appropriate fee structure, although he does not derive cardholder and merchant demand from first principles. For a non-extreme case, he shows the privately and socially optimal interchange fee coincide. Rochet and Tirole [2002a] relax assumptions (1) and (2), providing an explicit model of why competing merchants accept cards, and in so doing, incorporate the market interaction between consumers and merchants that arises.[4] By deriving consumer demand from first principles, Rochet and Tirole are able to consider the full welfare effects of different interchange fees, allowing for the effects on cash-paying consumers as well. However, as Schmalensee [2002, p.106] notes, because Rochet and Tirole assume all merchants are identical, their model cannot capture the trade-off between cardholder demand and merchant demand which the model of Schmalensee identifies.

This paper provides a tractable model of a card association relaxing assumptions (1)–(3) simultaneously. It incorporates the balancing role of interchange fees as in Schmalensee, as well as taking into account the effects of merchant competition as in Rochet and Tirole. The model is used to address the policymakers' concerns over interchange fees, and provides two distinct reasons why interchange fees can be set too high (or too low), relating these to the findings of the existing literature.

The primary role of the interchange fee is to determine the fee structure for cards. A higher interchange fee raises merchants' fees and lowers card fees. To maximize the volume of card transactions, the interchange fee is set so as to balance the extra card usage from a higher interchange fee against the loss in card transactions from lower merchant acceptance, in order to maximize the product of consumer and merchant demand for cards. The privately optimal interchange fee coincides with this output maximizing level, except to the extent acquirers pass-through interchange fee costs to merchants at a greater rate than issuers rebate interchange revenue to card users. Banks, in aggregate, will only sacrifice output by increasing the interchange fee above the level which maximizes the total number of card transactions if by doing so, they can increase merchant fees more than any decrease in card fees (or increase in card usage rebates).

The welfare maximizing interchange fee involves a trade-off between getting the right price signal for consumers and getting the right price signal for merchants. The right price for merchants is such that they accept cards whenever the sum of their transactional benefits from accepting cards and the average transactional benefits of their card-paying customers exceeds joint costs. The right price for consumers is such that they use cards

[4] Rochet and Tirole model merchants as competing according to the standard Hotelling model. Wright [2003a] considers the extreme cases of Bertrand competition between merchants and of monopoly merchants.

whenever the sum of their transactional benefits from using cards and the average transactional benefits from merchants they purchase from with cards exceeds joint costs. Except for very special cases, such as those considered in Wright [2003b], these goals are conflicting and a single instrument (the interchange fee) cannot achieve both conditions. Assuming issuers and acquirers pass through costs at the same rate, any divergence between the profit and welfare maximizing interchange fees only depends on a rather obscure asymmetry – an asymmetry in the difference in transactional benefits between inframarginal and marginal users across the two sides of the system.

Using specific models of merchant competition, stronger results are obtained. When merchants compete according to Hotelling competition within each industry, consumers are fully informed of whether merchants accept cards, and issuers and acquirers pass through costs at the same rate, any divergence between the profit and welfare maximizing interchange fees turns on whether the average transactional benefits merchants obtain from accepting cards are higher or lower than the merchant fee, or equivalently, whether on average, cash customers pay more for goods because others pay with cards. This model is used to highlight the impact of the business-stealing motive for merchants to accept cards, which implies a higher profit and welfare maximizing interchange fee than otherwise would be the case. The case in which merchants have no strategic motive for card acceptance is also considered. For a non-trivial set of cases, as Schmalensee found, the output, profit and welfare maximizing interchange fee are identical.

The rest of the paper proceeds as follows. Section II details the model, which is first analyzed at a general level in Section III and then analyzed for specific models of merchant competition in Section IV. Policy implications are discussed in Section V, including a discussion of the rule card schemes use to prevent merchants adding a surcharge to card purchases. Section VI briefly concludes.

II. MODEL SET-UP

The model is based most closely on the model of Rochet and Tirole. The main difference is the allowance for heterogenous merchants. Other key differences are that acquirers are allowed to be imperfectly competitive, and the card fee (which can be negative to represent rebates and other loyalty points) is set per transaction rather than per card. Wherever possible, the same notation is used. Here, the main features of Rochet and Tirole's framework are briefly reviewed, and the differences highlighted.

A transaction that is done using cards costs the issuing banks c_I and the acquiring banks c_A. Acquirers pay issuers a fee a, the interchange fee, for each card transaction. Given net per-transaction costs of $c_I - a$, competing symmetric issuers set an equilibrium per-transaction fee to card users of f

(which can be negative to reflect the various financial benefits given to card users such as cash rebates, reward points, and interest-free benefits). Similarly, competing symmetric acquirers, facing net per-transaction costs of $c_A + a$, set an equilibrium per-transaction fee to merchants of m. These equilibrium fees are functions of the interchange fee. We define markups per-transaction as the difference between per-transaction fees and per-transaction costs, with these being

$$\pi_I(a) = f(a) + a - c_I \tag{1}$$

and

$$\pi_A(a) = m(a) - c_A - a. \tag{2}$$

For simplicity, costs and fees that arise per cardholder or per merchant (rather than per transaction) are assumed away. In most countries, there are no membership or annual fees for debit cards. For credit cards, annual fees for cardholders are present in many countries, although options without such fixed fees are often available. For instance, in the U.S., 63 percent of issuers do not charge fixed fees (Chakravorti and Shah, 2003, p.8). Merchants generally face low or no fixed fees for accepting cards. Wright [2001] considers the extension of the model to fixed cardholder and merchant costs and fees.

A continuum (measure one) of separate industries and of consumers is assumed. Consumers are exogenously matched with industries, and so, without loss of generality, each consumer is matched with all industries. To abstract from the substitutability of goods between industries, consumers are assumed to wish to purchase one good from each industry they are matched to. Consumers get utility v from each such purchase. This approach captures the idea that consumers do not choose which industry to purchase from based on whether merchants in that industry accept cards or not even though, as will become clear, within the industry consumers may select the merchant to buy from based on whether the particular merchant accepts cards or not.

All merchants in an industry get the same benefit from accepting cards. When a merchant (seller) accepts a card for a sale rather than cash, she gets a net convenience benefit of b_S, where b_S is drawn independently for each industry from the common distribution function $G(b_S)$. This distribution has a positive and continuously differentiable density $g(b_S)$ over its support $[\underline{b}_S, \overline{b}_S]$. Thus, different industries are distinguished by the different benefits merchants obtain from accepting cards for payment versus the alternative cash (and so are referred to by their type b_S). This assumption is motivated by the fact that the costs of handling cash (and the convenience of accepting cards) differs primarily across different industries rather than across firms within an industry.

For given fees f and m, the critical value of b_S (the point at which merchants are indifferent between accepting cards or not) is denoted $b_S^m(f, m)$, so that for all industries with $b_S \geqslant b_S^m(f, m)$, merchants in these industries will accept cards and for all industries with $b_S < b_S^m(f, m)$, merchants will reject cards. In Section IV(i) we show that such a $b_S^m(f, m)$ can be defined for the Hotelling model of merchant competition, and in Section IV(ii) we do likewise for the case in which merchants do not accept cards for any strategic purpose. The measure of merchants that want to accept cards (or the supply of merchants that accept cards) is defined as $S(b_S^m) = 1 - G(b_S^m)$, and the average transactional benefit to those merchants accepting cards is the increasing function

$$\beta_S(b_S^m) = E\left[b_S \mid b_S \geqslant b_S^m\right] = \frac{\int_{b_S^m}^{\overline{b}_S} b_S g(b_S) db_S}{1 - G(b_S^m)} \text{ for } b_S^m < \overline{b}_S . \tag{3}$$

When a consumer (buyer) uses a card for a purchase, rather than cash, he is assumed to get a net convenience benefit of b_B, where b_B is drawn for each consumer from the common distribution function $H(b_B)$.[5] This distribution has a positive and continuously differentiable density $h(b_B)$ over its support $[\underline{b}_B, \overline{b}_B]$. Equivalently, each consumer could face the same expected convenience benefit of using cards, with his specific draw of b_B differing with each transaction.

For a given interchange fee, the critical value of b_B (the point at which consumers are indifferent between using a card or not) is denoted $b_B^m(f, m)$, so that all consumers with $b_B \geqslant b_B^m(f, m)$ will want to use cards and all consumers with $b_B < b_B^m(f, m)$ will not want to use cards. Since we assume merchants are not able to price-discriminate between the two types of payments or across consumers[6], consumers will use cards at a merchant that accepts cards whenever their transactional benefit b_B exceeds the fee f. Thus, $b_B^m = f$ for all cases we look at, and since f is a function of the interchange fee, b_B^m can be written as a function of a directly. The demand for card usage by consumers is defined as $D(b_B^m) = 1 - H(b_B^m)$, and the average transactional

[5] Following the existing literature which models interchange fees, the paper does not explicitly address the credit functionality available in some payment cards. Chakravorti and To [2000] and Chakravorti and Emmons [2003] provide models which focus on the credit aspect of payment systems, but neither paper explicitly models the optimal interchange fee.

[6] This can be justified by the presence of the no-discrimination rule set by card associations, or by the observation of price coherence even when merchants are free to surcharge – see Frankel [1988]. If such a condition does not apply to any merchant, interchange fees will be neutral, as Gans and King [2003] demonstrate in a general setting.

benefit to those consumers using cards is the increasing function

$$
\beta_B(b_B^m) = E\left[b_B \mid b_B \geqslant b_B^m\right] \tag{4}
$$
$$
= \frac{\int_{b_B^m}^{\overline{b}_B} b_B h(b_B) db_B}{1 - H(b_B^m)} \text{ for } b_B^m < \overline{b}_B .
$$

We make the following assumptions on these functions:

(A1) There is an interval $[\underline{a}_B, \overline{a}_B]$, such that $b_B^m(\underline{a}_B) = \overline{b}_B$, $b_B^m(\overline{a}_B) = \underline{b}_B$, and f and b_B^m are continuously differentiable with $df/da < 0$ and $db_B^m/da < 0$ over the interval.

(A2) There is an interval $[\underline{a}_S, \overline{a}_S]$, such that $b_S^m(f(\underline{a}_S), m(\underline{a}_S)) = \underline{b}_S$, $b_S^m(f(\overline{a}_S), m(\overline{a}_S)) = \overline{b}_S$, and m and b_S^m are continuously differentiable with $dm/da > 0$ and $db_S^m/da > 0$ over the interval.

(A3) The following parameter restrictions hold $\underline{a}_S \leqslant \underline{a}_B < \overline{a}_S \leqslant \overline{a}_B$.

(A4) $\pi_I(a) > 0$ and $\pi_A(a) > 0$ over the interval $[\underline{a}_B, \overline{a}_S]$.

(A5) There exists an interchange fee in the interval $[\underline{a}_B, \overline{a}_S]$ such that $\beta_B(b_B^m) + \beta_S(b_S^m) > c_I + c_A$.

Assumption (A1) says, over the range of interchange fees for which some (but not all) consumers want to use cards, an increase in the interchange fee lowers card fees and increases the proportion of consumers using cards. This will be true for both models of merchant behavior we consider. Similarly, (A2) says, over the range of interchange fees for which some (but not all) merchants want to accept cards, an increase in the interchange fee raises merchant fees and decreases the proportion of merchants accepting cards. This is consistent with the models of merchant behavior considered.[7] Assumption (A3) is then the simplest way to rule out the possibility that the privately or socially optimal interchange fee involves a corner solution (at a point where either all consumers or all merchants use cards). Assumption (A4) ensures that over the range of relevant interchange fees, symmetric issuers and symmetric acquirers set equilibrium fees that exceed their costs. Finally, (A5) ensures there exists at least one interchange fee for which card transactions can deliver positive welfare.

The timing of the game is summarized as follows:

(i) The payment card association or policy maker sets the level of the interchange fee a.

(ii) Competing issuers and acquirers set equilibrium fees $f(a)$ and $m(a)$.

[7] Even in the case of Hotelling competition where consumers are fully informed, a sufficient condition is that the rate at which issuers rebate interchange revenues is not too much higher than the rate at which acquirers pass through interchange costs.

(iii) Based on their individual realizations of b_B and b_S, consumers and merchants decide whether to make use of the payment network for purchases.

Section IV expands on what happens in (iii) for specific models of merchant behavior – explaining how retail prices are determined and how consumers choose which merchants to purchase from.

III. OPTIMAL INTERCHANGE FEES

In this section, the framework developed above is used to characterize the level of the privately and socially optimal interchange fees, explaining how these relate to the interchange fee which maximizes the volume of card transactions. The starting point to understand socially optimal interchange fees is to note, as Baxter does, that the first-best solution involves card transactions occuring if and only if the sum of transactional benefits exceeds the sum of transactional costs ($b_B + b_S \geqslant c_I + c_A$).[8] Given heterogenous consumers and merchants, this requires far more power and information on the part of the planner or the network operator than can be reasonably assumed. It requires that the planner can directly control which particular transactions are made using cards and which are not.

If, more feasibly, the planner can only influence (through the control of the fees charged to consumers and merchants) which consumers will use the payment card and which merchants will accept the payment card, the planner will generally want to select fees that imply Baxter's condition $b_B + b_S \geqslant c_I + c_A$ is violated for some transactions. In particular, a welfare maximizing social planner would pick fees f and m so as to maximize total welfare[9]

$$W(f,m) = \int_{b_B^m(f,m)}^{\overline{b}_B} \int_{b_S^m(f,m)}^{\overline{b}_S} (b_B + b_S - c_I - c_A) g(b_S) h(b_B) db_S db_B \tag{5}$$

$$= (\beta_B(b_B^m) + \beta_S(b_S^m) - c_I - c_A) D(b_B^m) S(b_S^m) \tag{6}$$

The second-best solution calls for some transactions to be made for which the joint benefit to the marginal card user and marginal merchant will be less than the total cost of the transaction, since even such marginal card users provide a positive contribution to social surplus once the contribution from transactions with other merchants that accept cards is taken into account (and likewise for the marginal merchant that accepts cards). This suggests there are likely to be too few card transactions given that banks will set fees

[8] As Schmalensee [2002] has noted, this assumes the alternatives to using cards (cash and checks) are themselves efficiently provided.

[9] Note that in constructing total welfare, transfers between cash-paying customers, card-paying customers, merchants, issuers, and acquirers cancel out due to the assumption of unit demands.

to at least cover their costs, and an outside subsidy to the card system can potentially increase welfare. However, this need not be the case. The business stealing effect, in which merchants accept cards to take business from rivals, can lead merchants to accept cards even when their transactional benefits are less than the fee they face, and it is possible for there to be too much card acceptance.

A card association does not get to select the fees charged to consumers and merchants. Instead, it sets the level of the interchange fee. When an interchange fee is used to try to optimize the size of the network, it can have at best only a limited effect. Since an interchange fee is a transfer from one side of the system to the other, it cannot be used to change the overall number of transactions other than by altering the balance between card usage and merchant acceptance. Put differentially, for a card association where card fees and merchant fees are set in a decentralized way, a single interchange fee cannot simultaneously achieve the optimal level of both fees. Rather, the main effect of a single interchange fee will be to get the right structure of fees between consumers and merchants, while the overall level of fees (card fee plus merchant fee) will be pinned down by competition between members of the card association, and between different payment systems. In determining the structure of card fees versus merchant fees, the optimal interchange fee will normally involve a trade-off between promoting card usage and merchant acceptance.

Implicit in the above discussion is the assumption that the card association cannot set different interchange fees that apply to different classes of consumer and merchant transactions. The assumption of a single interchange fee is consistent with our timing assumption, which implies the transactional benefits that particular consumers and merchants get from using cards are unobserved by the card scheme when setting the interchange fee. If the card scheme can observe information on the transactional benefits that merchants in particular industries obtain (or elicit such information through a mechanism such as that explained by Crémer and McLean, 1985), the card scheme (or a central planner) may be able to do better by setting multiple interchange fees that apply to different industries. Although there is some limited use of different interchange fees across different industries (for example, there is generally a lower interchange fee for transactions at supermarkets in the United States), card fees do not usually reflect these differences (consumers do not pay higher transaction fees, or earn lower rebates when they spend at a supermarket versus other retailers) so the analysis of such a setting will differ both from the analysis in this paper and that in Rochet and Tirole's standard model.[10] We leave an analysis of

[10] Rochet and Tirole note how their model can be modified for this case, but an analysis of the socially optimal interchange fee is not conducted.

multiple interchange fees in such a setting for future research and instead focus on the case in which schemes either do not observe the transactional benefits that consumers and merchants obtain, or where they do, still only set a single interchange fee.

In setting a single interchange fee, card schemes face a trade-off which we examined at a general level in this section. A useful benchmark that enables the profit and welfare maximizing interchange fees to be compared is the interchange fee which maximizes the volume of card transactions, which is considered first.

III(i). *Output Maximizing Interchange Fee*

The interchange fee which maximizes the volume of card transactions is denoted a^T. The total number of card transactions is

$$T(b_B^m, b_S^m) = \int_{b_B^m}^{\overline{b}_B} \int_{b_S^m}^{\overline{b}_S} g(b_S)h(b_B)db_S db_B \tag{7}$$

$$= D(b_B^m)S(b_S^m). \tag{8}$$

Assumptions (A1)–(A3) ensure that T is continuously differentiable with respect to a, has a positive value within $[\underline{a}_B, \overline{a}_S]$, and equals zero everywhere else. The assumptions ensure that a global maximum of T must exist for an interchange fee between $\underline{a}_B$ and $\overline{a}_S$. The first order condition for output maximization is

$$\begin{aligned} \frac{dT}{da} &= S\frac{dD}{da} + D\frac{dS}{da} \\ &= 0, \end{aligned} \tag{9}$$

one of the solutions of which will signal the global maximum.[11] For simplicity, it is assumed this stationary point is unique in $[\underline{a}_B, \overline{a}_S]$, so it is indeed the global maximum. At the margin, the output maximizing interchange fee balances the increase in consumer demand for cards resulting from lower card fees (this has to be multiplied by the proportion of merchants that accept cards to obtain the impact on total demand) with the decrease in merchant demand for accepting cards resulting from higher merchant fees (this has to be multiplied by the proportion of consumers that use cards to obtain the impact on total demand).

[11] Rochet and Tirole [2002b] obtain a similar result for a single proprietary card scheme that sets f and m directly to maximize its profit. This underscores the similarity in fee structures across proprietary and joint-venture forms.

III(ii). *Profit Maximizing Interchange Fee*

The card association maximizes the profit of its members, who collectively decide its interchange fee.[12]

$$\Pi(b_B^m, b_S^m) = \int_{b_B^m}^{\bar{b}_B} \int_{b_S^m}^{\bar{b}_S} (\pi_I + \pi_A) g(b_S) h(b_B) db_S db_B = (\pi_I + \pi_A) T(b_B^m, b_S^m), \tag{10}$$

and the interchange fee which maximizes (10) is denoted a^{Π}, the profit maximizing interchange fee. Given (A4), $\Pi > 0$ in the interior of $[\underline{a}_B, \overline{a}_S]$. Assumptions (A1)–(A3) ensure $\Pi = 0$ elsewhere, and that Π is continuously differentiable with respect to a over $[\underline{a}_B, \overline{a}_S]$. Therefore, a global maximum of Π must exist for an interchange fee between $\underline{a}_B$ and $\overline{a}_S$.

The joint profit maximizing interchange fee involves a trade-off between maximizing the total number of card transactions and maximizing profit per transaction. The first-order condition from maximizing (10) with respect to a is

$$\frac{d\Pi}{da} = (\pi_I + \pi_A)\left(S\frac{dD}{da} + D\frac{dS}{da}\right) + \left(\frac{d\pi_I}{da} + \frac{d\pi_A}{da}\right)T = 0. \tag{11}$$

For simplicity, it is assumed this stationary point is unique over $[\underline{a}_B, \overline{a}_S]$, so it corresponds to the global maximum. For given joint profits per transaction $(\pi_I + \pi_A)$, the first line in (11) is the condition for output maximization from (9). For a given volume of card transactions (T), the second line is the condition for maximizing joint profit *per transaction.*

Proposition 1. The privately optimal interchange fee is higher (lower) than the output maximizing interchange fee if and only if the pass through of costs to user fees is higher (lower) on the acquiring side than the issuing side, when evaluated at the output maximizing interchange fee.

[12] Schmalensee [2002] considers the possibility that issuers will have greater bargaining power than acquirers. In our model, this possibility can be captured by weighting π_I by more than π_A. The implication is that the card association's optimal interchange fee will be higher than the output maximizing level if the pass-through of costs by issuers and acquirers is equal and less than one, and will be lower than the output maximizing level if the pass-through of costs by issuers and acquirers is equal and greater than one. Only if banks pass through costs one-for-one, so that their markups do not depend on the interchange fee, will the weight given to issuer versus acquirer profits not affect results. Otherwise, the interchange fee is set away from the output maximizing level in order to increase the equilibrium profits of the side of the market which has greater control over setting interchange fees.

Proof. The output maximizing interchange fee satisfies $S(dD/da)+D(dS/da)=0$. Substituting this result into (11) implies

$$\frac{d\Pi}{da}=\left(\frac{d\pi_I}{da}+\frac{d\pi_A}{da}\right)DS,$$

so $d\Pi/da \gtrless 0$ at $a=a^T$ if and only if $d\pi_I/da \gtrless -d\pi_A/da$ at $a=a^T$ or equivalently if and only if $dm/dc_A \gtrless df/dc_I$ at $a=a^T$. □

Whenever higher interchange fees increase per-transaction profits to issuers more than they decrease per-transaction profits to acquirers, the expression in the second line of (11) will be positive. The profit maximizing interchange fee will be higher than the interchange fee which maximizes output, with some transactions being sacrificed in order to transfer per-transaction profits to the side of the market where they will be competed away less. Alternatively, if costs are passed through by the same amount on both sides of the market, then the output and profit maximizing interchange fee will coincide.

III(iii). *Welfare Maximizing Interchange Fee*

The interchange fee which maximizes total welfare in (5) is denoted a^W. Assumptions (A1)–(A3) ensure that W is continuously differentiable with respect to a over $[\underline{a}_B, \overline{a}_S]$, and equals zero for $a \leqslant \underline{a}_B$ and $a \geqslant \overline{a}_S$. Assumption (A5) ensures that W takes on at least one positive value in $[\underline{a}_B, \overline{a}_S]$. It follows a global maximum of W must exist for an interchange fee between $\underline{a}_B$ and $\overline{a}_S$. The first-order condition for maximizing (5) with respect to a is

$$\begin{aligned}\frac{dW}{da} &= S\frac{dD}{da}(b_B^m+\beta_S(b_S^m)-c_I-c_A)\\ &\quad + D\frac{dS}{da}(\beta_B(b_B^m)+b_S^m-c_I-c_A)\\ &= 0.\end{aligned} \tag{12}$$

For simplicity, it is assumed this stationary point is unique over $[\underline{a}_B, \overline{a}_S]$ and so (12) characterizes the global maximum.

To interpret the welfare maximizing interchange fee, consider a small increase in the interchange fee. As a result of lower card fees there will be some additional consumers who will now want to use cards for transactions whereas previously they did not (this is measured by dD/da). The increase in surplus arising from the additional consumers who now want to use cards depends on the number of merchants that accept cards (S) multiplied by the social benefits averaged over these additional transactions (this equals $b_B^m+\beta_S(b_S^m)-c_I-c_A$). On the other hand, as a result of higher merchant fees, there will be some industries where merchants no longer want to accept cards even though previously they did (this is measured by dS/da). The decrease in surplus arising from this fall in merchant acceptance depends on the proportion of consumers that would have wanted to use cards at these merchants (D) multiplied by the social benefits that would have arisen when averaged over these transactions (this equals $\beta_B(b_B^m) + b_S^m - c_I - c_A$).

Welfare is maximized by setting a fee structure so that as many transactions where joint transactional benefits ($b_B + b_S$) exceed joint costs ($c_I + c_A$) take place using cards, and as many transactions where $b_B + b_S < c_I + c_A$ take place using cash. Using a single interchange fee, this objective is best achieved if merchants only accept cards when the sum of their transactional benefit from accepting cards (b_S) and the average transactional benefits of their card-paying customers ($\beta_B(b_B^m)$) exceed joint costs, and if consumers only use cards when the sum of their transactional benefit from using cards (b_B) and the average transactional benefits from merchants they purchase from with cards ($\beta_S(b_S^m)$) exceed joint costs. This requires an interchange fee such that both $b_B^m + \beta_S(b_S^m) = c_I + c_A$ and $b_S^m + \beta_B(b_B^m) = c_I + c_A$. In general, both conditions cannot simultaneously be satisfied. Instead, welfare maximation is achieved by setting an interchange fee which balances the impact on consumers' decisions to use cards and the surplus such card usage creates with the impact on merchants' decisions to accept cards and the surplus such merchant acceptance creates.

Clearly, if $b_B^m + \beta_S(b_S^m) = b_S^m + \beta_B(b_B^m)$ at the output maximizing interchange fee, then (12) coincides with (9), and the welfare maximizing and output maximizing interchange fees are identical. Otherwise, it follows that

Proposition 2. The welfare maximizing interchange fee is higher (lower) than the output maximizing interchange fee if and only if the transactional benefits of the marginal card user plus the average transaction benefit of merchants who accept cards is higher (lower) than the transactional benefits of the marginal merchant who accepts cards plus the average transactional benefit of consumers who use cards, when evaluated at the output maximizing interchange fee.

Proof. The output maximizing interchange fee a^T satisfies $S(dD/da) + D(dS/da) = 0$. Substituting this result into (12) implies

$$\frac{dW}{da} = \left[(b_B^m + \beta_S(b_S^m)) - (b_S^m + \beta_B(b_B^m))\right] \frac{dD}{da} S,$$

so given (A1), $dW/da \gtreqless 0$ at $a = a^T$ if and only if $b_B^m + \beta_S(b_S^m) \gtreqless b_S^m + \beta_B(b_B^m)$ at $a = a^T$. □

The result has a natural interpretation. If the condition $b_B^m + \beta_S(b_S^m) > b_S^m + \beta_B(b_B^m)$ holds at the output maximizing interchange fee, by increasing the interchange fee above the level which maximizes output, there will be fewer card transactions, but the gain in surplus for the marginal card user who now uses cards and all those merchants who accept his cards will be greater than the loss in surplus from the marginal merchant who no longer accepts cards and all those card users who can no longer use their cards for purchases at her store. It is an asymmetry in the difference between the benefits received by inframarginal and marginal users across card users and merchants that drives any divergence between the output and welfare maximizing interchange fee.

III(vi). *Private Versus Social Optimum*

Section III(ii) showed that the output maximizing interchange fee is the same as the profit maximizing interchange fee except to the extent that there is an asymmetry in the pass-through of costs between issuers and acquirers, in which case interchange fees can be used to raise profits by restricting the volume of card transactions. Section III(iii) showed that the output maximizing interchange fee is the same as the welfare maximizing interchange fee except to the extent that there is an asymmetry in the marginal and infra-marginal benefits of card users and merchants, in which case interchange fees can be used to raise welfare by restricting the volume of card transactions.

Combining the results in propositions 1 and 2 implies that in general, the profit maximizing interchange fee could be higher or lower than the welfare maximizing interchange fee, and could involve more or fewer card transactions when compared to the welfare maximizing level. A case in which the profit maximizing interchange fee exceeds the welfare maximizing interchange fee and involves excessive card transactions (the policymakers' case), corresponds to when issuers and acquirers pass through costs at the same rate, but at the profit maximizing interchange fee the additional benefits of inframarginal merchants over and above the marginal merchant accepting cards is less than the additional benefits of inframarginal card users over and above the benefits of the marginal card user.

Alternatively, the profit maximizing interchange fee is higher than the welfare maximizing interchange fee but there are too few card card transactions if the main source of the divergence between a^{Π} and a^{W} is that acquirers pass on interchange costs into merchant fees at a greater rate than issuers compete away interchange revenues. In this case, regulation that aims to reduce credit card transactions will be counter productive. There are also cases in which the profit maximizing interchange fee is less than the welfare maximizing interchange fee and there are either too many card transactions or too few card transactions.

Rochet and Tirole [2002a] find that without merchant heterogeneity, the welfare maximizing interchange fee is always less than or equal to the profit maximizing level, and that there can never be under-usage of cards. In their setting, issuers' profit always increases in the interchange fee provided merchants accept cards, as does card usage, while perfectly competitive acquirers make zero profit anyway. This implies the profit maximizing and the output maximizing interchange fee coincide. Moreover, the welfare maximizing interchange fee involves setting the right price signal to consumers to use cards so they internalize the merchants' benefits from accepting cards, subject of course to the constraint that merchants still accept cards at this interchange fee.

The introduction of merchant heterogeneity gives rise to two new effects relative to Rochet and Tirole's analysis of interchange fees. First, with perfectly competitive acquirers and imperfectly competitive issuers, proposition 1 implies that the card scheme will face a trade-off – profit maximization will involve the card scheme trading off higher issuer *margins* through higher interchange fees with the reduction in output from the decrease in merchant demand. In contrast to Rochet and Tirole's analysis, this can result in too *few* card transactions at the profit maximizing interchange fee. Second, the welfare maximizing interchange fee now involves a balancing act – trying to get both consumers and merchants to face the right price signals since the decision of each type of user now impacts on the number of card transactions and the surplus that is created from card transactions. This differs from the interchange fee that is optimal in Rochet and Tirole, and in general there is nothing stopping it being higher than the profit maximizing interchange fee.

IV. SPECIFIC MODELS OF CARDHOLDER/MERCHANT DEMAND

Up to this point, consumer and merchant behavior has been left quite general. To interpret the above results further, some more structure is put on consumer and merchant demand. In Section IV(i), we apply Rochet and Tirole's model of merchant competition to each industry in our framework. This provides a generalization of Rochet and Tirole's model to heterogenous merchants. In Section IV(ii), we consider the case in which merchants do not accept cards for strategic reasons. Examples include a special case of the Hotelling model in which consumers are not informed of which merchants accept cards, as well as the case in which all merchants are monopolists. This provides rigorous foundations for Schmalensee's results. In both cases we assume that in stage (iii) of the game, merchants decide whether to accept cards or not; in stage (iv), they set their retail prices; and in stage (v) consumers decide which merchants to buy from and whether to use cards or not.

IV(i). *Hotelling Competition Between Merchants*

Each industry is assumed to be made up of two merchants, who compete according to a Hotelling model of competition. Consumers are randomly located in each industry according to the standard 'linear city' version of the Hotelling model, and the two merchants are located at the two extremes of the unit interval. Consumers draw an x for each industry from the $U[0, 1]$ distribution, and incur transportation costs of tx if they purchase from firm 1 and $t(1-x)$ if they purchase from firm 2. This draw is independent of their draw of b_B. Like Rochet and Tirole (section 5.2), we assume that before making their choice of which merchant to purchase from, consumers observe whether merchants in a particular industry accepts cards or not with probability α.

Given the assumption of no price discrimination, and since consumers face no membership fee, consumers will use cards whenever the transactional benefits of doing so b_B exceed the fee f. Thus, consumers use cards whenever

$$(13) \qquad b_B \geqslant b_B^m = f.$$

By accepting cards, merchants get transactional benefits b_S but pay a merchant fee m. If accepting cards attracts no additional customers, this is all that matters. This happens $1-\alpha$ of the time. For the remaining α of the time, when merchants accept cards they are able to attract additional customers from rivals who do not accept cards. In such cases, by accepting cards, a merchant gets net transactional benefits of $b_S - m$ and is able to offer more to its customers by allowing them to obtain the surplus from using cards, which on average equals $\beta_B(f) - f$ (it can capture this surplus in higher margins or additional sales). Thus, competing merchants will accept cards whenever $(1-\alpha)(b_S - m) + \alpha(b_S - m + \beta_B(f) - f) > 0$ or

$$(14) \qquad b_S \geqslant b_S^m = m - \alpha(\beta_B(f) - f).$$

An appendix, posted on the Journal's editorial web site, demonstrates that (14) in fact defines an equilibrium in which both merchants accept cards, provided $b_S \geqslant b_S^m$, while when $b_S < b_S^m$ it is a unique equilibrium for both merchants to reject cards. This is consistent with the definition of b_S^m in Section II.

Combining these conditions with propositions 1 and 2, we get

Proposition 3. When merchants compete according to Hotelling competition, consumers are fully informed of which merchants accept cards ($\alpha = 1$) and issuers and acquirers pass through costs at the same rate, the welfare maximizing interchange fee will be higher (lower) than the profit maximizing interchange fee if and only if at the profit maximizing interchange fee the average transactional benefit over all those merchants who accept cards is higher (lower) than the fee they pay.

Proof. From proposition 2, $a^W \gtrless a^T$ if and only if $b_B^m + \beta_S(b_S^m) \gtrless b_S^m + \beta_B(b_B^m)$ at $a = a^T$. From (13) and (14) this is equivalent to the condition that $\beta_S(b_S^m) \gtrless m + (1-\alpha)(\beta_B(b_B^m) - f)$ or $\beta_S(b_S^m) \gtrless m$ given $\alpha = 1$. Given the symmetry of issuers and acquirers, $a^T = a^\Pi$ from proposition 1 so that $a^W \gtrless a^\Pi$ if and only if $\beta_S(b_S^m) \gtrless m$. □

The result gives a relatively simple condition under which the profit maximizing interchange fee is either too high or too low compared to the social optimum. It says that if at the privately set interchange fee the average benefit that merchants obtain from accepting cards (over all merchants that choose to accept cards) exceeds the merchant fee they face, then welfare can be increased by forcing the card association to increase its interchange fee (which will reduce card transactions). The reverse result holds when average merchant benefits fall short of the fee they face, in which case a lower interchange fee raises welfare and lowers card transactions. As Rochet and Tirole found, it is possible to have too many card transactions, but these results show to correct for this does not necessarily require a decrease in the interchange fee to maximize welfare.

The condition in proposition 3 turns out to be closely related to whether average retail prices increase or decrease as a result of card acceptance. The equilibrium retail price set by two merchants in an industry of type b_S is $d + t + D(f)(m - b_S)$ if they accept cards and $d + t$ if they do not. Contrary to the case with homogeneous merchants, the effect of card acceptance on a merchant's retail prices is ambiguous. The average retail price across merchants who accept cards is

$$\bar{p} = d + t + D(f)(m - \beta_S(b_S^m)). \tag{15}$$

Where merchants accept cards for strategic reasons (that is, to attract additional business), there will be some merchants who accept cards even though the transactional benefits they obtain are less than the merchant fees they pay. For these merchants, retail prices will be higher than if they did not accept cards. Equation (15) implies that averaged across all card accepting merchants, retail prices can be higher or lower than for those merchants who do not accept cards, and so the question of whether cash-paying customers pay more as a result of the existence of card-paying customers is an empirical one. Comparing equation (15) to the condition in proposition 3, we get

Proposition 4. When merchants compete according to Hotelling competition, consumers are fully informed of which merchants accept cards ($\alpha = 1$) and issuers and acquirers pass through costs at the same rate, average retail prices for merchants accepting cards will be higher (lower) as a result of the existence of cards if and only if the welfare maximizing interchange fee is higher (lower) than the profit maximizing interchange fee.

Proposition 4 demonstrates that there is a close link between efficiency arguments that interchange fees are set too high (or too low) and equity arguments that consumers who pay by cash are made worse off (better off) by the existence of cards. Proposition 4 shows that one cannot presume, as Frankel [1988] does, that cash-paying customers necessarily pay more as a result of the existence of more expensive card-paying customers – one has to consider the additional benefits the cards provide as well.

The above two propositions provide useful characterizations of whether the welfare maximizing interchange fee is higher or lower than the profit maximizing interchange fee since they depend on simple and potentially measurable conditions. A more insightful, though perhaps less measurable characterization of the difference between a^W and a^Π can be obtained by noting the formulation of welfare at equation (6). Then the welfare maximizing interchange fee satisfies the alternative first order condition

$$(16) \qquad \frac{dW}{da} = (\beta_B + \beta_S - c_I - c_A)\frac{dT}{da} + \left(\frac{d}{da}(\beta_B + \beta_S)\right)T = 0.$$

The welfare maximizing interchange fee thus involves a trade-off between the average surplus per transaction and the total number of card transactions.

Provided issuers and acquirers pass through costs at the same rate so that $a^T = a^\Pi$, the welfare maximizing interchange fee will only deviate from the profit maximizing interchange fee if by charging more to one side of the system (thereby reducing the number of card transactions), the average cardholder and merchant benefits per card transaction can be increased. Lowering the interchange fee below the one that maximizes profit, increases card fees and decreases merchant fees. When merchants accept cards for strategic reasons (to attract customers), competing merchants internalize their customers' benefits from using cards. In this case, the effect of a decrease in merchant fees is largely offset by the effect of an increase in card fees on merchants' decision, about whether to accept cards or not. The main effect will then be that fewer consumers will use cards, resulting in a higher average transactional benefit from card usage. This suggests that the welfare maximizing interchange fee involves some sacrifice of card transactions in order to raise the average transactional benefit from card usage, which is achieved with an interchange fee below the profit maximizing one.

The result holds for instance in the case b_B and b_S are distributed according to the uniform distribution. Given the assumption that issuers and acquirers pass through costs at the same rate, the uniform distributions imply

$$\frac{d}{da}(\beta_B + \beta_S) = \frac{d}{da}\left(\frac{f + \bar{b}_B}{2} + \frac{m - \alpha\left(\frac{\bar{b}_B - f}{2}\right) + \bar{b}_S}{2}\right)$$
$$= \frac{\alpha}{4}\frac{df}{da} < 0$$

provided $\alpha > 0$. It follows from (16) that $a^W < a^\Pi$.

If we assume, in addition, that issuers and acquirers set a constant markup per transaction[13], then closed form solutions for a^Π and a^W can be obtained. That is, assume

(17) $$f = c_I - a + \bar{\pi}_I$$

and

(18) $$m = c_A + a + \bar{\pi}_A.$$

Then the privately optimal interchange fee is

(19) $$a^\Pi = \frac{(\bar{b}_S - c_A - \bar{\pi}_A) - (1 - \alpha)(\bar{b}_B - c_I - \bar{\pi}_I)}{(2 - \alpha)}$$

and the socially optimal interchange fee is

(20) $$a^W = a^\Pi + \frac{1}{3\alpha(2 - \alpha)}\Bigg((4 - 3\alpha)\Delta + (4 - \alpha)\rho - 2\sqrt{(4 - 6\alpha + 3\alpha^2)\Delta^2 + 8(1 - \alpha)\Delta\rho + (4 - 2\alpha + \alpha^2)\rho^2}\Bigg)$$

where $\rho = \bar{\pi}_I + \bar{\pi}_A > 0$ and $\Delta = \bar{b}_B + \bar{b}_S - c_I - c_A > 0$.

An important determinant of optimal interchange fees apparent in the formulas above is the extent to which merchants accept cards for strategic reasons. This is measured here by the value of α. Differentiating (19) with respect to α and using (A5), it is straightforward to show that in the limit as $\bar{\pi}_I = \bar{\pi}_A \to 0$, the card association's preferred interchange fee is increasing in α. When merchants accept cards for strategic reasons, this both increases the number of merchants who will want to accept cards and makes the number of merchants accepting cards less sensitive to changes in the interchange fee. Other things equal, this means that, starting from charging both consumers and merchants equal fees, increasing the fee to merchants and decreasing the fee to consumers will increase the total volume of card transactions, and

[13] For instance, this arises if issuers are identical Bertrand competitors, as are acquirers, but they can only set retail fees in discrete units.

hence the card association's profit. By the same logic, the strategic motive for accepting cards also increases the welfare maximizing interchange fee. This can be shown by differentiating (20) and considering the limit case in which $\bar{\pi}_I = \bar{\pi}_A \to 0$. However, since the card association does not take into account the impact of attracting additional card usage on inframarginal users, in this case, it sets an interchange fee that is too high.[14]

The case is clearly only a specific one. More generally, whether the socially optimal interchange fee is higher or lower than the privately set fee will also depend on distributional assumptions on b_B and b_S, as well as any asymmetry in pass-though of costs between issuers and acquirers, and the particular reasons merchants accept cards. The next section examines the case in which there is no strategic reason for accepting cards.

IV(ii). *No Strategic Motive to Accept Cards*

The above model of merchant demand was based on merchants' accepting cards, in part, to attract customers from each other. In the special case in which consumers do not observe whether merchants accept cards before they choose which merchant to purchase from ($\alpha = 0$), this strategic role for accepting cards disappears. Merchants will only accept cards if the transactional benefits of accepting cards exceed the merchant fee so that $b_S^m = m$. This is Baxter's assumption but here we can extend Baxter's setting to the case with merchant heterogeneity. As Wright [2003a] shows, this merchant acceptance condition also applies to the case in which there is a separate monopolist in each industry. These cases provide a first-principles justification for the model of Schmalensee since in these cases merchants' willingness to pay is a measure of the transactional benefits they receive. Welfare will therefore correspond to the Marshallian surplus measure which Schmalensee uses to evaluate optimal interchange fees.

To the extent firms compete, they will pass on the transactional benefits they get from accepting cards, and so prices will be lower as a result of the acceptance of payment cards. Using the Hotelling model of competition above but where $\alpha = 0$, the average retail price across merchants that accept cards will be $\bar{p} = d + t + D(f)(m - \beta_S(m))$, which is necessarily lower than the price without cards (since $\beta_S(m) > m$). Cash-paying customers unambiguously benefit from the existence of card-paying customers.

Like Schmalensee, we can consider a linearized version of this model. In particular, suppose b_B and b_S are both distributed according to the uniform distribution so that $D(f)$ and $S(m)$ are linear functions. Provided the equilibrium fees f and m are linear in the interchange fee and that issuers and

[14] This follows since $(4 - 6\alpha + 3\alpha^2)\Delta^2 + 8(1 - \alpha)\Delta\rho + \left(4 - 2\alpha + \alpha^2\right)\rho^2 > \left(\frac{1}{2}((4 - 3\alpha)\Delta + (4 - \alpha)\rho)\right)^2$ which is true provided $(\Delta - \rho)^2 \alpha^2 > 0$; that is, $\alpha > 0$ and $\Delta \neq \rho$.

acquirers pass through costs at the same rate, so that

$$f = r(c_I - a) + \bar{\pi}_I \tag{21}$$

and

$$m = r(c_A + a) + \bar{\pi}_A, \tag{22}$$

then demands will be linear in the interchange fee and it can be shown that

$$a^W = a^\Pi = a^T = \frac{c_I - c_A}{2} + \frac{\bar{\pi}_I - \bar{\pi}_A + \bar{b}_S - \bar{b}_B}{2r}. \tag{23}$$

Like Schmalensee, we obtain the result that for linear demand, the profit maximizing interchange fee also maximizes total output and welfare.[15]

The higher the benefits to merchants of accepting cards (relative to the benefits to consumers from using cards), and the lower the costs of (and margins on) servicing merchants (relative to servicing card users), the more merchants will accept cards relative to the proportion of consumers who use cards. Other things being equal, this implies the to maximize the volume of card transactions, one needs more card users and fewer merchants who accept cards, which can be achieved via a higher interchange fee. Given the symmetry between issuers' and acquirers' pass-through of costs, proposition 1 implies the profit maximizing interchange fee will also maximize the volume of card transactions. To understand why the welfare maximizing interchange fee also involves maximizing the volume of card transactions, recall the condition for welfare maximization in equation (16). At the output (and profit) maximizing interchange fee, the first term in (16) is zero, while the second term is also zero since with linear demand $\beta_B(f) = (f + \bar{b}_B)/2$ and $\beta_S(m) = (m + \bar{b}_S)/2$. In this case, there is no way to restrict output in order to raise the average transactional benefits across card users and merchants who accept cards since the sum of such benefits is independent of the interchange fee, reflecting the symmetry between the two sides of the system.

V. POLICY IMPLICATIONS

In this section, two aspects of policy are explored in light of the above analysis. First, does the theory provide any justification for the policy-makers' proposed regulation of interchange fees? Second, the implications of the model for the no-surcharge rule used by card schemes are noted.

[15] Schmalensee considers the case where there is a single monopoly issuer and a single monopoly acquirer, which is consistent with (21) and (22) when $r = 1/2$.

V(i). *Regulation of the Interchange Fee*

The Reserve Bank of Australia recently moved to regulate interchange fees of credit card associations based on a component of their issuers' costs (see RBA, 2002). The regulations are aimed at achieving substantially lower interchange fees in Australia. Frankel [1988] has gone further, proposing interchange fees be regulated to zero. Neither cost-based nor zero interchange fees finds any support from the analysis in this paper. Even in the special case of linear demands and almost perfect competition, the socially optimal interchange fee in (20) is complex and depends on cardholder and merchant demands, as well as cost factors. Cost-based or zero interchange fees will be optimal only by chance.

The above analysis does permit the possibility that privately set interchange fees will be too high. It also shows that it is possible for interchange fees to be set too low, that there may be too many card transactions, or too few. To disentangle the different possibilities, additional empirical evidence is needed. Policymakers and proponents of interchange fee regulation have not provided any such evidence. One aspect on which empirical evidence could shed light is on the possibility of an asymmetry in pass through of costs between issuers and acquirers. According to the theory above, if issuers pass through costs less than acquirers, then the interchange fee may be set too high in order to restrict output and raise members' profit. Even in the presence of such asymmetries, the extent of any such restriction may be minor as Schmalensee's discussion of alternative system structures demonstrates. This is reinforced by the existence of inter-system competition. In any case, no evidence that interchange fees are being used to restrict output has been put forward by proponents of regulation; in fact, the complaint is usually that there is excessive card use.

The policymaker's claim that interchange fees are set too high must therefore be based on a failure of card schemes to properly internalise the effects of their fee structure on inframarginal merchants who accept cards, resulting in interchange fees being set too high. In this respect, card schemes are no different from almost any other commercial enterprise. We do not normally expect businesses to take into account effects on inframarginal users in setting their price structures. Normally, it is sufficient that businesses do not engage in practices that restrict output to raise prices. Society accepts there will sometimes be excessive entry by firms or that retailers will engage in excessive advertising. In part, this reflects the fact that the information required to identify and appropriately deal with these kinds of distortions is usually far too demanding for policymakers. It is notable that even when one makes the strongest simplifying assumptions in our model, the socially optimal interchange fee that takes these inframarginal effects into account depends on a complicated combination of cost, demand and profit factors.

The presence of inter-system competition is only likely to magnify this complexity.

A further reason to caution against any regulation of interchange fees arises from considering the impact of competition from proprietary schemes. Regulating a lower interchange fee within card associations so as to restrict the excessive use of their payment cards will provide proprietary schemes with a competitive advantage. Since proprietary schemes set card fees and merchant fees directly, they do not need to use an interchange fee to set their optimal fee structure. Interchange fee regulation thus amounts to imposing an unpopular fee structure on MasterCard and Visa, while leaving proprietary schemes such as AMEX free to set their own fee structures. There is no reason to suppose that proprietary schemes have an incentive to match this unpopular fee structure and thereby disadvantage themselves. Instead, proprietary schemes are likely to maintain their present fee structures, attracting business away from card associations, even if they have higher costs or are otherwise less desirable.

V(ii). *The No-Surcharge Rule*

An underlying assumption in the modelling is that consumers face the same price for cash or card purchases. If merchants are free to surcharge and discount in the model, then it follows from Gans and King [2003] that the level of the interchange fee will be neutral. The level of the interchange fee will not affect the decision of merchants to accept cards, the choice of payment instrument by consumers, the banks' profit or the number of card transactions. Gans and King show this is true regardless of the type of merchant competition, or the extent of heterogeneity. However, absent rules preventing surcharging or discounting, survey evidence from the Netherlands and Sweden suggests that in practice, the vast majority of merchants choose not to discount or surcharge at all (see IMA Market Development AB, 2000 and ITM Research, 2000). Provided some merchants do not discount or surcharge over the range of interchange fees under consideration (and the selection of these merchants is not itself determined by the level of the interchange fee), then the analysis in this paper is still applicable even if the no-surcharge rule is relaxed.

Assuming away the frictions that cause most merchants to set a single price for cash and card purchases, one can also use the model to address the welfare implications of various rules used by card schemes to prevent merchants' surcharging. For instance, under the no-surcharge rule, merchants are free to discount for cash. Strictly interpreted, the model implies the no-surcharge rule will have no effect given there are only two alternative payment instruments and a discount for one is equivalent to a surcharge for the other. The implications of a stronger rule, such as the no price discrimination rule that prevents firms setting differential prices between cash and card purchases, will depend on the nature of merchant behavior.

Where merchants compete imperfectly, as represented by the Hotelling model in Section IV(i), the welfare effects of the rule are, in general, ambiguous as Rochet and Tirole [2002a] show. For this case, the same trade-off they uncover still applies here. Relaxing the rule is desirable to the extent it allows consumers and merchants to use cards whenever there is a joint benefit to them from doing so. Merchant heterogeneity makes this effect more important as merchants will be able to offer different surcharges and discounts to reflect their particular transactional benefits of accepting cards. On the other hand, since under surcharging consumers will face all the costs of the card network (including covering the issuers' and acquirers' margins), but only some of the benefits, there will be too little card usage which can be corrected by imposing the no-discrimination rule and setting the interchange fee appropriately.

In the case in Section IV(ii) where merchants are monopolies, merchants will be able to use surcharging to extract additional surplus from card users, and the no-surcharge rule will tend to be desirable (Wright, 2003a).[16] Alternatively, where merchants are perfectly competitive, they will anyway separate into those that accept cards and those that do not, and the rule will be irrelevant since interchange fees will be neutral in such an environment (Wright, 2003a).

VI. CONCLUSIONS

This paper has provided a simple theoretical framework to analyze the determinants of interchange fees in payment card associations. It builds on the existing work of Baxter [1983], Rochet and Tirole [2002a], and Schmalensee [2002] by taking into account heterogeneity of both consumers and merchants, and imperfect competition between issuers and between acquirers. Unlike the findings of Rochet and Tirole's model in which merchants are homogenous and acquirers perfectly competitive, the socially optimal interchange fee in this paper involves a trade-off between getting consumers to face the right price signal to use cards and merchants to face the right price signal to accept cards. In general, the model demonstrates that the privately set interchange fee can be higher or lower than the socially optimal one, and can involve more or fewer card transactions.

One version of the model, in which merchants do not accept cards for strategic reasons, corresponds to Schmalensee's framework, thus providing a first principles justification for his partial demands and Marshallian

[16] See also Schwartz and Vincent [2002] who also consider the case in which merchants have monopoly power, although in their setting, the no-surcharge rule can either increase or decrease total welfare.

surplus approach. In this case, when partial demand functions are linear, the output, profit and welfare maximizing interchange fees all coincide. Alternatively, when merchants accept cards to attract business from each other, as in Rochet and Tirole's model, the optimal interchange fee was found to be higher.

The paper highlights two sources of deviation between the privately and socially optimal interchange fees. Privately optimal interchange fees may be too high if merchant fees increase with interchange fees but issuers do not rebate the additional interchange fee revenue back to cardholders. In this case, high interchange fees are a way to transfer profits to the side of the scheme where they are least competed away, resulting in a restriction in output. On the other hand, socially optimal interchange fees may be higher or lower than the profit maximizing interchange fee because of an asymmetry in inframarginal effects. This reflects the fact that the usage decision of each type of user affects the transactional benefits obtained by inframarginal users of the opposite type. If there is any asymmetry in these inframarginal effects, the fee structure should reflect this, something that may not be taken into account in the scheme's private choice of interchange fee. We showed how this source of market failure can be linked to the issue of whether retail prices are higher or lower as a result of card acceptance. Although the model highlighted the theoretical possibility of a deviation between the privately and socially optimal interchange fees, it also highlighted the gap between the the arguments being put forward by proponents of interchange fee regulation and a sound basis for any such regulation.

REFERENCES

Baxter, W. F., 1983, 'Bank Interchange of Transactional Paper: Legal Perspectives,' *Journal of Law and Economics*, 26, pp. 541–588.

Chakravorti, S. and Emmons, W. R., 2003, 'Who Pays for Credit Cards?,' *Journal of Consumer Affairs*, forthcoming.

Chakravorti, S. and Shah, A., 2003, 'Underlying Incentives in Credit Card Networks,' *The Antitrust Bulletin*, forthcoming.

Chakravorti, S. and To, T., 2000, 'A Theory of Credit Cards,' mimeo, Federal Reserve Bank of Chicago.

Chang, H. H. and Evans, D. S., 2000, 'The Competitive Effects of the Collective Setting of Interchange Fees by Payment Card Systems,' *The Antitrust Bulletin*, 45, pp. 641–677.

Crémer, J. and MacLean, R. P., 1985, 'Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist When Demands are Interdependent,' *Econometrica*, 53, pp. 345–362.

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

Frankel, A. S., 1998, 'Monopoly and Competition in the Supply and Exchange of Money,' *Antitrust Law Journal*, 66, pp. 313–361.

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

IMA Market Development AB, 2000, 'Study Regarding the Effect of the Abolition of the Non-Discrimination Rule in Sweden for Europian Commission Competition Directorate General,' February.

ITM Research, 2000, 'The Abolition of the Non-Discrimination Rule,' Amsterdam, March.

Lyons, D., 2002, 'Visa's Vision,' *Forbes Magazine*, September 4.

RBA., 2002, 'Reform of Credit Card Schemes in Australia. IV. Final Reforms and Regulation Impact Statement,' Reserve Bank of Australia, public document, August.

Rochet, J-C. and Tirole, J., 2002a, 'Cooperation Among Competitors: Some Economics of Credit Card Associations,' *Rand Journal of Economics*, 33, pp. 549–570.

Rochet, J-C. and Tirole, J., 2002b, 'Platform Competition in Two-Sided Markets,' mimeo, December 13, IDEI and GREMAQ.

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

Schwartz, M. and Vincent, D., 2002, 'Same Price, Cash or Card: Vertical Control by Payment Networks,' Working Paper 02-01 Georgetown University.

Wright, J., 2000, 'An Economic Analysis of a Card Payment Network,' mimeo, NECG and University of Auckland.

Wright, J., 2001, 'Determinants of Optimal Interchange Fees in Payment Systems,' University of Auckland Working Paper No. 220.

Wright, J., 2003a, 'Optimal Card Payment Systems,' *European Economic Review*, 47, pp. 587–612.

Wright, J., 2003b, 'Pricing in Debit and Credit Card Schemes,' *Economic Letters*, 80, pp. 305–309.