KONKURRENSVERKETS
WORKING PAPER SERIES
IN LAW AND ECONOMICS

WORKING PAPER 2015:1

Economics of Payment Cards

Ozlem Bedre-Defolie and Linda Gratz

KONKURRENSVERKET
Swedish Competition Authority

Konkurrensverket Working Paper series in Law and Economics covers current research topics in the field of competition and public procurement policy that may be of interest to a wider public.

Opinions expressed are those of the author(s) and do not necessarily represent the views of Konkurrensverket.

Konkurrensverket Working Paper 2015:1
Author: Ozlem Bedre-Defolie and Linda Gratz
Konkurrensverket, September 2015

Economics of Payment Cards

Özlem Bedre-Defolie[a] and Linda Gratz[b]

March 2015

Abstract

This article summarizes the literature on two-sided payment card markets. The general conclusion is that interchange fees can help internalize the complementarity between services on both sides of the market but private platforms set too high interchange fees from a social welfare perspective. Private platforms' price structure is distored in favor of buyers for several reasons: asymmetric choices between buyers and merchants, merchant internalization and/or platform competition. Furthermore, market power of platforms leads to higher total user prices similar to the case of one-sided markets. Platform competition can help to correct for such market power distortions but may exacerbate the price structure distortions.

It is shown that full efficiency in the industry cannot be achieved through regulating the interchange fee. This is because the interchange fee affects only the allocation of the total user price between buyers and sellers. The first-best efficiency also requires a lower total price level due to positive externalities between the two sides. A measure to test whether interchange fees are excessive has been proposed but this measure may be imperfect.

JEL Classification: L11, G21, L42, L51, K21.

Keywords: payment card networks, interchange fees, two-sided markets

[a]European School of Management and Technology - Berlin. Email: ozlem.bedre@esmt.org.
[b]E.CA Economics - Berlin. Email: gratz@e-ca.com.

1 Introduction

In 2013, around 44% of all payments in the EU and around 54% of all payments in the US were made by card.[1] For each card transaction, the merchant has to pay a merchant fee to its bank and in most card networks the merchant's bank pays an interchange fee to the cardholder's bank. In these networks interchange fees constitute the major part of merchant fees. In the UK, for example, IFs make up around 70% of the merchant fees.[2] Even though interchange fees are rather small on a single transaction level (in the US, 1.8% and in the EU, between 0.1% and 1.5%) merchants annually spend approximately $3.26 trillion in the US and €1.8 trillion in the EU on credit, debit or prepaid cards.[3] In contrast, cardholders are often offered rewards if they checkout by card. Hence, there seems to be an asymmetry between the costs and benefits of card payments on the side of cardholders versus merchants. This asymmetry is mainly due to agreements among banks within card payment networks such as MasterCard and Visa, in which banks multilaterally decide on interchange fees. Antitrust authorities and regulators are concerned that the asymmetric pricing policies by banks restrict competition between merchants' banks and inflate merchants' costs of card acceptance without enhancing the efficiency of the system. A debate has been triggered on whether interchange fees should be capped, and if so, at what level.

This paper summarizes important aspects of the payment card industry that distinguish its analysis from standard markets and reviews the literature documenting the main sources of market failures in this industry. We focus on market inefficiencies driven from the payment card platform's short-run pricing decisions, that is, when optimal fee choices do not correspond to the socially optimal levels maximizing the total social welfare. The paper then also briefly discusses potential market interventions that the policy makers could implement to address these failures.

1.1 Background of the payment card industry

Antitrust investigations and regulations in the payment card market target in particular Interchange Fees (IFs). In four-party payment card networks, such as Visa and MasterCard, these are fees

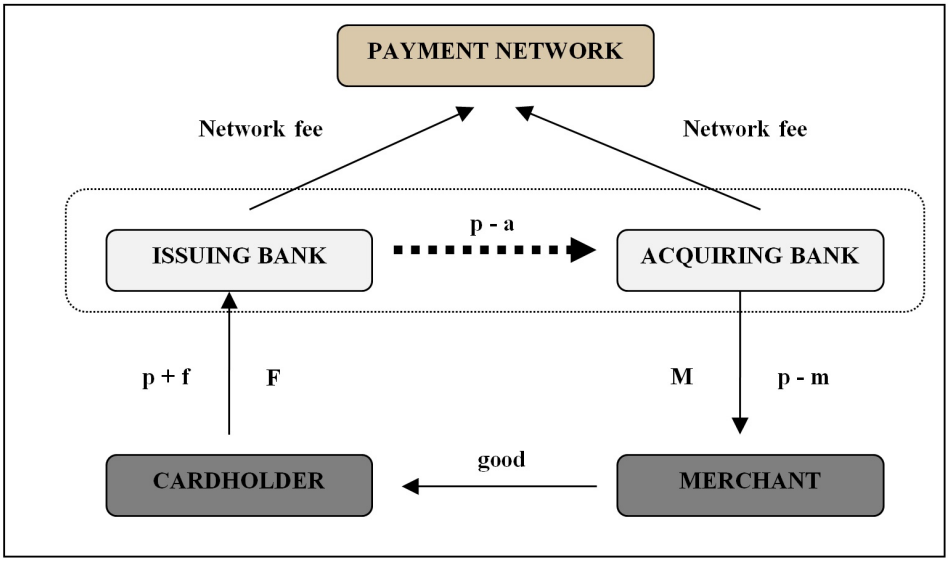
[1] See press release of the European Central Bank (ECB) from 9 September 2014 on "Payment statistics for 2013", Table 1, and Nilson Report, Issue 1054, Dec 2014, chart "Consumer Payment Systems in the U.S. 2013 vs. 2018".

[2] See OFT Report (July 2012).

[3] See Visa Europe decision (February 2014, p. 18), Nilson Report (July 2009, issue 929, pp. 1), Nilson Report (April 2007, issue 895, pp. 7) and Hayashi (2009).

multilaterally agreed by member banks, charged by a cardholder's bank (issuer) to a merchant's bank (acquirer) for each sales transaction made at a merchant outlet with a payment card. Figure 1 illustrates the fees in a four-party card network where the issuer remits the transaction amount (p) less the IF (a) to the acquirer, keeping the IF.

Figure 1: Four-party payment card network



Note: p denotes the price of the good, F denotes the card membership fee, f denotes the card transaction fee, M denotes the merchant membership fee, m denotes the merchant fee per transaction and a denotes the multilateral interchange fee.

In card networks, like in Figure 1, when a cardholder uses a payment card to buy from a merchant, the acquirer pays the merchant the sales price (p) less a merchant fee (m), a fee that the merchant has to pay to its bank for accepting the card as a means of payment. The issuer enters the entire transaction amount (p) and in some cases also a card transaction fee (f) to the cardholder's account.

The role of issuers is to provide services such as connectivity to the card network, terminal hardware and software, and customer support. They have to make investments in innovation, security, efficient payments administration, etc. Issuers take greater risks and costs than acquirers who only transmit IFs and electronic authorization requests. Evidence shows that the profitability of issuing is higher than acquiring in the EU and in the US.[4]

A payment card network can also involve three instead of four parties. In this case, a single payment network provider serves as both issuer and acquirer. Examples of three-party payment

[4]See e.g. Brenning-Louko, Panova, Repa, and Teixeira (2006).

schemes are AMEX, Diners Club, and Discover.

The level of merchant fees might depend on the merchant and business sector, and on the type of card and transaction. IFs are higher for credit cards than debit cards, for international networks than domestic ones. For example in the EU, "MIF levels show wide divergence between Member States. Regarding consumer card transactions, their weighted average level ranges from between 0.1-0.2% to 1.4-1.5% in the Member States."[5]

The markets where platforms create value by connecting two (or more) distinct groups of customers and facilitating interactions between them by lowering transaction costs and search costs are called two-sided markets.[6] The payment card industry is a two-sided market where payment networks facilitate interaction, in this case card transactions, between two groups of users, here, merchants and cardholders. A card transaction requires the participation of these two different groups of users.

Platforms can use end-user prices as a tool to balance demands of both sides. The theoretical literature of two-sided markets has documented that optimal monopoly platform pricing involves charging a lower price to the side whose demand is relatively more sensitive to prices (high elasticity of demand) and charging a lower price to the side that generates relatively more value to the other side (Rochet and Tirole, 2006). In four-party card networks IFs determine the allocation of the total end user prices between merchants and cardholders, given that acquirers pass-through IFs (at least partially) to merchants by raising merchant fees and issuers pass-through IFs (at least partially) to consumers by lowering card transaction fees. In three-party networks the networks contract directly with users and so determine merchant fees and card transaction fees directly.

Two-sided markets are distinct from standard markets due to the existence of "network" externalities between the different groups of users, that is, the value of participating in a platform for one group of users depends directly on the amount of participation of the other group. For example, the more consumers hold cards of a network, say Visa, the more merchants are willing to accept Visa cards. The more merchants accept Visa cards, the more consumers would like to hold Visa cards. Studies of markets with network effects started much earlier than the literature on two-sided markets. Two features distinguish two-sided markets from standard markets with network effects.

[5] See Visa Europe decision (February 2014).

[6] The literature on two-sided markets was pioneered by Caillaud and Jullien (2003), Rochet and Tirole (2003, 2006), Armstrong (2006) and, more recently, Weyl (2010).

First, they exhibit indirect network effects between different groups of users. Second, platforms can price discriminate between these groups.

Rochet and Tirole (2006) distinguish network externalities from "usage" externalities. A platform market exhibits usage externalities if one side benefits/loses from a usage decision made by the other side of the market. In some markets, among others in the payment card market, both types of externalities exist. In the payment card market, however, consumers decide on both network membership and card usage, while merchants only decide on membership. That is, in the payment card market there are one-sided usage externalities from cardholders to affiliated merchants. Every time cardholders choose to pay by card, merchants have to pay a merchant fee to their bank and might enjoy the convenience benefits of being paid by card.

Gans and King (2003) show that the corresponding externalities between the two sides could in principle be internalized if merchants were able to price discriminate based on the payment method without frictions. If merchants and cardholders were to bargain over usage choices on a purchase-by-purchase basis, they would internalize the usage externalities through side payments. The effect of IFs on card transactions would be neutral, independent of the degree of competition at either the bank or the merchant level.

However, in practice it is uncommon for sellers to price discriminate based on the payment method, mainly due to two reasons. First, (even very small) transaction costs can make it unprofitable to price discriminate. Second, price discrimination is often prohibited by payment networks' rules, for example, the No Surcharge Rule (NSR) which bans surcharging payments by using the cards of the network against any other payment method, which might include cash, possibly distorting competition. As a result, users of a payment platform cannot perfectly internalize those externalities and so the volume of transactions depends on the allocation of the total transaction fees $(f + m)$ between the two sides, and thus on the level of IFs. Therefore, the analysis of pricing in the payment card industry differs from the standard theory of taxation, where it does not matter whether the tax is on sellers or buyers. Unlike firms selling complementary products to the same set of customers, platforms sell complementary services that are consumed by two distinct groups of users exerting externalities on each other (Rochet and Tirole, 2003).

Antitrust authorities are concerned that the asymmetric pricing policies by banks, which allocate large parts of the total end user prices on the merchants' side by setting relatively high

IFs, inflate merchants' costs of card acceptance without enhancing the efficiency of the system. A debate has been triggered as to whether IFs should be capped, and if so, at what level. In general, card payments are associated with significant welfare benefits. For cardholders the benefits of card versus cash payments are greater convenience, forgoing the costs of cash withdrawals or converting foreign currency, speed, and security. Merchants benefit from lower costs of cash holdings, greater security, faster payments, and the processing of incoming transactions. With regard to the applicable welfare standard it must be distinguished between consumer surplus, that is, the welfare of cardholders, total user surplus, that is, the welfare of cardholders and merchants, and social welfare, that is, the welfare of cardholders, merchants, and banks/platform providers. The social optimum is reached when the cardholders make the efficient decision with regards to the choice of payment method.

1.2 Brief overview of the theoretical literature

Most of the literature focuses on understanding the optimal pricing incentives of card payment platforms and the banks within these platforms, whether these incentives differ from those of a social planner, and if so, on the direction of efficiency distortion resulting from privately optimal prices. A common finding is that due to its market power a monopoly platform upwardly distorts the total level of per-transaction prices (i.e. $f+m$ in Figure 1). Furthermore, a monopoly platform distorts the price structure that is the allocation of the total price between the two sides. Platform competition would correct the market power distortion on the total price level, at least partially. However, the direction of the price structure distortion is not straightforward, and so it remains unclear whether platform competition also reduces the distortion of the price structure (see, for instance, Rochet and Tirole, 2003, 2006; Weyl, 2010).

For the payment card industry, the latter result implies that the privately set interchange fees (and so the merchant fees) might be inefficiently higher or lower than the socially optimal level. A similar conclusion is driven from the earlier literature on interchange fees (Wright, 2004; Schmalensee, 2002).[7]

Recent work has documented reasons why payment platforms' profitable pricing strategies

[7]For a detailed overview of this literature see Evans and Schmalensee (2005), Verdier (2011), and Chakravorti (2010).

would lead to inefficiently high interchange fees. Rochet and Tirole (2002) and Wright (2012) show that card networks set inefficiently high merchant fees when merchants accept cost-increasing cards as a way to steal customers from their rivals. Rochet and Tirole (2003), Guthrie and Wright (2007), and Armstrong (2006) illustrate that if sellers accept the cards of multiple card networks (multi-home), competition increases the distortion of the price structure even further as networks try to woo cardholders back from their rivals by lowering their card fees. Furthermore, Bedre-Defolie and Calvano (2013) explain yet another source of bias against sellers by the fact that once sellers have decided whether to accept payment cards and buyers have decided whether to hold cards, buyers are the ones who decide on the extent of card usage.

Most theoretical models on the pricing of payment cards assume that sellers cannot or do not surcharge card payments. This assumption seems to hold in practice as either card networks impose the no-surcharge rule (NSR) or sellers hardly make use of the possibility to surcharge. From a theoretical point of view, the effect of the NSR on social welfare is ambiguous.

A measure to test whether interchange fees are excessive has been proposed by Rochet and Tirole (2011) and applied by many policy makers thereafter. Under this test an interchange fee (which induces a certain merchant fee) passes the test if and only if accepting a card payment does not increase the sellers' operating cost. However, this is only an exact test of excessive interchange fees from the point of view of total user surplus when certain conditions are met.

1.3 Structure of the paper

The paper is structured as follows. In section 2, we describe theoretical arguments for potential reasons of market failures in the payment card industry. Section 4 discusses the welfare effects of the NSR. Section 3 summarizes the current state of the regulation of interchange fees. We conclude in section 5.

2 Theory of pricing payment cards

A first economic defense of IFs was suggested by Baxter (1983) during the course of the *NaBanco* litigation.[8] Baxter argues that IFs enable four-party networks to internalize the network exter-

[8]NaBanco is the most prominent antitrust challenge to credit card interchange fees in the US. The court held that Visa's IF was necessary to the Visa system, achieved efficiencies, and could be avoided by using another network.

nalities between consumers and merchants. Baxter shows that the socially optimal IF is equal to the merchants' transaction benefit minus the acquirers' costs ($a^* = b_S - c_A$) under two important assumptions : 1) Merchants do not differ in their transaction benefit from card and any other payment method. 2) Acquirers and issuers are perfectly competitive, with acquirers charging $m^*(a) = c_A + a$ as merchant fees. Given "Baxter's IF" (a^*), the equilibrium merchant fee equals merchants' transaction benefit, $m^*(a) = c_A + a^* = b_S$, and the equilibrium transaction fee for cardholders is equal to the opportunity cost of attracting one more cardholder, that is, the transaction cost of issuing minus the net benefit that one more cardholder generates on the merchant side: $f^*(a) = c_I - a^* = c_I - (b_S - c_A)$. Hence, in equilibrium consumers internalize their externality on the merchants' side and an efficient volume is induced as consumers use cards whenever $b_B + b_S >= c_I + c_A$, where b_B denotes cardholders' transaction benefit.

Absent an IF, issuers and acquirers would not internalize the complementarity of their services fully and would set card and merchant fees above the levels prevailing with an IF. In other words, IFs would enable issuers and acquirers to apportion the aggregate prices for their services between them in the proportions represented by the height of their respective demand curves at the output level, at which their aggregate benefits are maximized.

If issuing and acquiring banks are not perfectly competitive, IFs enable four-party networks to avoid the double mark-up problem that arises due to the fact that two complementary services, acquiring and issuing, are sold independently. An economic defense of IFs is that they can lead to an internalization of externalities and avoid the double mark-up problem, creating efficiencies within four-party networks.

The subsequent theoretical literature on the pricing of payment cards reveals several potential reasons for market failures. First of all, Rochet and Tirole (2003, 2006) predict that if a platform has market power, the total user fees will be too high compared to the social optimum. Further, given the two-sided market structure, payment card networks must choose not only a price level but also a price structure for their services. Rochet and Tirole (2003, 2006) predict that a monopoly platform will distort the allocation of the total user fees between the two sides due to the platform's inability to price discriminate across heterogeneous users on each side. In a model that distinguishes

The court took the view that Visa lacked market power as it applied a broad market definition which included all forms of payments.

between extensive and intensive margins Bedre-Defolie and Calvano (2013) identify the direction of this distortion. They show that interchange fees are set too high compared to the social optimum due to the different decisions of buyers and sellers: buyers decide on membership and usage while sellers only decide on membership. This leads to a distortion in a monopoly platform's pricing structure in favor of buyers because the equilibrium price allocation over-subsidizes the side which determines the usage volume for a given membership level and over-taxes the other side.

Rochet and Tirole (2003) and Guthrie and Wright (2007) show that a platform's pricing structure is also distorted if two networks compete and consumers single-home more than sellers. In fact, competition increases the distortion of the allocation of the total user price even further because networks will try to woo cardholders back from their rivals by lowering their prices. Networks can then charge sellers the monopoly price to provide access to their exclusive turf of cardholders. If both buyers as well as sellers multi-home, the privately optimal interchange fees are higher than in case of a monopoly network (Bedre-Defolie and Calvano, 2013).

Merchant internalization is shown to be another complementary source of the distortion of a platform's pricing structure. Rochet and Tirole (2002, 2011) point out that card networks set inefficiently high merchant fees when competing merchants accept cost-increasing cards as a way to steal customers from their rivals. The greater the competitive edge guaranteed by card acceptance, the easier it is to convince merchants to join the network, and the more likely it is that card networks will exploit the lower merchant "resistance" to fee increases by setting inefficiently high merchant fees. Wright (2012) extends the finding by Rochet and Tirole (2002, 2011) to the case where the merchant demand for card acceptance is elastic (unobserved merchant heterogeneity).

In this section, we first outline the model of Bedre-Defolie and Calvano (2013) (hereafter BC) to show in section 2.2 by means of this model that platform market power leads to higher total user prices and distorts the platform's pricing structure in favor of buyers. It follows in section 2.3 a derivation from the BC benchmark model of the results on the distortions of a platform's pricing structure if two networks compete. Finally, in section 2.4 we expand on the effect of merchant internalization, the third complementary source of the distortion of a platform's pricing structure.

2.1 Benchmark model of Bedre-Defolie and Calvano (2013)

BC analyze pricing incentives in an open (four-party) card network which provides card payment services to buyers through a monopoly issuer bank and to sellers through perfectly competitive acquirer banks. For each card transaction, the issuer incurs cost c_I and the acquirer incurs cost c_A. The total transaction cost is $c = c_I + c_A$. The card network requires the acquirer to pay an interchange fee a per transaction to the issuer.

This market structure captures the fact that the issuing side of the market is widely regarded as having strong market power, whereas the acquiring side is found to be highly competitive (see, for instance, Evans and Schmalensee, 2005; Rochet and Tirole, 2002, 2003; and the EC's report, 2007). Besides, this setup is formally equivalent to a closed (three-party) network since with zero margins on the acquiring side, the network's choice of a is controlled by the issuer and so the issuer acts as a single platform owner, charging merchants for card services given that competitive acquirers simply pass on interchange fees to merchants.

There is a continuum (mass one) of buyers and a continuum (mass one) of local monopoly sellers. Buyers are willing to purchase one unit of a good from each seller and the unit value from consumption is assumed to be the same across sellers.[9] Let $v > 0$ denote the value of a good purchased in cash, that is, the consumption value net of all cash-related transaction costs. A buyer gets $v - p$ from purchasing a unit good by cash at price p and the seller gets p from this purchase (since retailing costs are set to zero without loss of generality). Assume that there is a price coherence, that is, the price of a good is the same regardless of whether it is paid in cash or by card, and that sellers are not allowed to steer buyers toward their preferred method of payment.

Buyers get an additional payoff of $b_B - f$ when they pay by card rather than in cash, where b_B denotes the buyer benefit from a card transaction and f is the per-transaction fee. Cardholders also pay a fixed (membership) fee, F, to the issuer. Similarly, sellers get an additional payoff of $b_S - m$ when paid by card, where b_S denotes the seller convenience benefit from a card payment and m denotes the (per-transaction) merchant fee to be paid to the acquirer. Acquirers also charge a fixed fee, M, to merchants for card acceptance.

Buyers and sellers are assumed to be heterogeneous in their usage benefits from card payments

[9] This assumption is to focus on buyers' card usage choices and therefore allows us to abstract away from the effects of card prices on the consumption demand. See Wang (2010) and Shy and Wang (2011) for related issues.

in the following way. Buyer benefit b_B is distributed over interval $\left[\underline{b_B}, \overline{b_B}\right]$ with the cumulative distribution function (cdf) $G(b_B)$ and probability density function (pdf) $g(b_B)$. Similarly, seller benefit b_S is distributed over interval $\left[\underline{b_S}, \overline{b_S}\right]$ with cdf $K(b_S)$ and pdf $k(b_S)$. Assume that $\underline{b_S} < \overline{b_S}$, $\underline{b_B} < \overline{b_B}$, and that $G(.)$ and $K(.)$ satisfy the Increasing Hazard Rate Property (IHRP).[10] To guarantee an interior solution to the pricing problems it is assumed that $\underline{b_S} + \underline{b_B} < c < \overline{b_S} + \overline{b_B}$.

There is no sign restriction on benefits and fees, potentially allowing for negative benefits, that is, distaste/intrinsic costs of card transactions and negative fees, for example, reward schemes like cash back bonuses or frequent-flyer miles.

The timing of the game is as follows:

Timing

Stage i: The payment card network (alternatively a regulator) sets the interchange fee a.

Stage ii: After observing a, the issuer sets its card fees and each acquirer sets its merchant fees.[11]

Stage iii: Sellers decide whether to accept the payment card and which bank to patronize depending on their transaction benefits b_S. Simultaneously, buyers decide whether to hold a payment card and which bank to patronize.

Stage iv: Sellers set retail prices. Buyers decide whether to purchase depending on their transaction benefits b_B. Finally, cardholders decide whether to pay by card or in cash.

Compared to previous models this model has a broader parameter space (allowing for ex-ante uncertainty on usage benefits) and a broader action space (allowing for non-linear prices). The grounds for this model (and thus for its timing) come from the observation that not all cardholders pay by card at all sellers that accept cards, which implies that transaction benefits indeed differ across transactions. Therefore, consumers get the card in order to secure the option (or expected value) of paying by card whenever this happens to be convenient for a particular transaction.

It is assumed that the card network sets the IF to maximize the sum of the profits earned by its issuers and acquirers. This assumption aims to represent the real objectives of for-profit card

[10] The IHRP leads to *log-concavity* of demand functions (for cardholding, for card usage, and for card acceptance), which is sufficient for the second-order conditions of the optimization problems CB solve.

[11] In the three-party network interpretation, the first two stages pin down to one stage where the network sets card fees and merchant fees.

associations.[12] In principle, for-profit card organizations could charge their members non-linear membership fees and could thus internalize any incremental increase in their members' profits through fixed transfers.[13] In the analysis, this means defining the profit of the network as the total fee collected from members, which could be proxied by the total profit of its member banks, allowing the network to charge fixed fees as well as transaction fees to its members.

To simplify the benchmark analysis BC assume that v is sufficiently high so that sellers never find it profitable to exclude cash users by setting a price higher than v. This assumption rules out the case where sellers try to extract some of the surplus associated with card transactions by increasing their retail prices. Thus, monopoly sellers set $p = v$ regardless of whether they accept card payments or not.

Buyers pay by card if and only if $b_B \geq f$, so buyers' quasi demand of card usage (i.e., the demand of a cardholder at each merchant) is $D_B(f) = Pr(b_B \geq f) = 1 - G(f)$, which is decreasing in the per-transaction fee f. Sellers accept cards whenever $b_S \geq m$, so sellers' demand is $D_S(m) = Pr(b_S \geq m) = 1 - K(m)$, which is decreasing in the merchant fee m. The average buyer surplus from card transactions is defined as $v_B(f) \equiv E[b_B - f | b_B \geq f]$ and the average seller surplus from card transactions is defined as $v_S(m) \equiv E[b_S - m | b_S \geq m]$.

2.2 Platform market power

Rochet and Tirole (2003) analyze the case in which a monopoly platform charges buyers and sellers per-transaction fees, fixed card membership fees being zero. They show that the market power of platforms distorts the total user price, similar to one-sided markets, which is referred to as "market power distortion" by Weyl (2010). A standard monopoly markup, $f + m > c$, is set on the total price such that $\frac{f+m-c}{f+m} = \frac{1}{\eta_B + \eta_S}$, where $\eta_B = -\frac{fD'_B}{D_B}$ is the elasticity of buyers' card usage demand and $\eta_S = -\frac{mD'_S}{D_S}$ is the elasticity of sellers' card acceptance demand.[14]

[12] Visa and MasterCard used to be non-profit organizations, but in 2003 Visa and in 2006 MasterCard became for-profit organizations in Europe and their shares are jointly owned by their member banks. See the EC's report (2007a) and EU Commission's Prohibition Decision on MasterCard from 2007.

[13] Indeed, in almost all countries where the card associations operate, a significant portion of the operating revenues are typically concentrated among a handful of large issuers. According to many industry observers, the card networks set their terms to maximize issuer profits. For example, Rochet and Tirole (2002), Wang (2010), Shy and Wang (2011) assume that this is the case.

[14] This result is obtained in the general literature on two-sided markets (see Rochet and Tirole, 2003, 2006, Armstrong, 2006, and Weyl, 2010) as well as in the literature focusing on the payment card industry (see Guthrie and Wright, 2007, and Wright, 2004, 2012).

Result 1: In a monopolistic card payment network, if users are charged only transaction fees, the total price set to users is too high from a social planner's perspective.

Furthermore, Rochet and Tirole (2003) show that profit-seeking platforms in general distort the structure of the total price between the two sides ("Spence distortion" as named by Weyl, 2010). This is because, while allocating the total price between the two sides, the private platform cares about the marginal user whereas the social planner cares about the average user. The equilibrium allocation of the total price between the two sides (price structure) is formally characterized by:[15]

$$\text{Planner's optimal: } \frac{f}{m} = \frac{\eta_B}{\eta_S} \div \frac{v_B}{v_S}$$
$$\text{Platform's optimal: } \frac{f}{m} = \frac{\eta_B}{\eta_S}$$

The socially optimal allocation of the total price, $f + m = c$, is achieved when the relative user prices are equal to the ratio of the relative demand elasticities divided by the ratio of the relative average surpluses per transaction of buyers and sellers. The privately optimal allocation of the total price, on the other hand, is achieved when the relative user prices are equal to the ratio of the relative demand elasticities.

A common finding is that the existence and sign of "Spence distortion" is not straightforward since it depends on variables that are hardly measurable (Rochet and Tirole, 2003; Wright, 2004; Schmalensee, 2002). Assessing "Spence distortion" would require a significant amount of information, and in principle an optimal intervention could go in either direction. Platform competition, for instance, would correct the "market power distortion" on the total price level but not necessarily the distortion on the price structure. Weyl (2010) points out that "Spence distortion" mainly depends on the source of user heterogeneity and Armstrong (2006) finds that it does not exist when there is only membership heterogeneity.

Within BC's model the direction of "Spence distortion" can be shown: A platform with market power sets an inefficient price structure, over-subsidizing card usage and over-taxing sellers. The *sign* of the price structure distortion does not depend on fundamental costs and/or preference

[15] An analogous property holds for the optimal access charge between backbone or telecom operators where the access charge allocates the total cost between two groups of users (consumers and web sites in backbone networks, call receivers and call senders in telecommunication networks) (see Laffont et al., 2003).

attributes, only its *magnitude* does. To see this, consider BC's benchmark model (see section 2.1).

As buyers decide on card membership and card usage, the platform has to consider two margins on the buyers' side: extensive margin (how the platform's pricing influences card membership) and intensive margin (how the platform's pricing influences card usage). Card membership and card usage are determined by issuers' two-part tariffs to buyers, including a per-transaction fee f and a fixed fee F. If the platform decreases the per-transaction fee by Δ and increases the fixed fee by $\Delta D_S(m)D_B(f)$, buyers' demand for cardholding (i.e., the extensive margin) would be unaffected, whereas buyers' quasi-demand for card usage (i.e., the intensive margin) would increase since $D'_B(f) < 0$. BC show that shifting the fee is profitable for issuers as long as their per-transaction margin is positive. Thus, the issuer sets its transaction fee at the effective marginal cost, $f = c_I - a$, and captures the buyers' expected transaction surplus via a fixed fee, $F = v_B(f)\, D_B(f)\, D_S(m)$, where $v_B(f) \equiv E\left[b_B - f \mid b_B \geq f\right]$ is the buyers' average surplus from card usage. In other words, two-part tariffs which include a fixed fee F and a per-transaction fee f are useful on the buyer side.

Unlike buyers, sellers only decide on card membership not on card usage. Once sellers have become a member, they cannot reject the transaction demand from buyers. Thus, on the seller side there is only an extensive margin (how the platform's prices influence membership) and two-part tariffs which include a fixed fee M and per-transaction fee m are redundant. For a given amount of card usage by buyers, say N_B, sellers' demand for card acceptance depends only on the average merchant fee, $m + M/N_B$. Hence, if the platform decreases m by Δ and increases M by ΔN_B, sellers' demand will be unaffected. It follows that without loss of generality the fixed merchant fee (and sellers' membership benefits) can be set to zero. The perfectly competitive acquirers set the merchant fee at their effective marginal cost, $m = c_A + a$.

Accordingly, the private monopoly platform sets the IF that maximizes *buyers' surplus* from transactions, $v_B(f)\, D_B(f)\, D_S(m)$. It does so because through two-part tariffs it can internalize the buyers' surplus but not the sellers' surplus. Two-part tariffs enable the platform to capture the usage surplus of an average buyer, whereas on the seller side the platform cares only about the marginal seller. This leads to a distortion of the equilibrium fees in favor of buyers. Too high prices are charged to merchants.

Unlike the private monopoly platform, a social planner sets an interchange fee that maximizes

the sum of the average user surpluses, $[v_B(f) + v_S(m)]\, D_B(f) D_S(m)$, subject to the banks' optimal fee choices, $f + m = c$, and so chooses the optimal allocation of the total user price between buyers and sellers.[16] That is, compared to the private monopoly platform it does not ignore the sellers' surplus from card transactions. Since at equilibrium prices the buyers' average surplus increases in IFs and the sellers' average surplus decreases in IFs, BC obtain the following result.

Result 2: Consider BC's setup of a monopolistic card payment network with a monopoly issuer and perfectly competitive acquirers where banks are allowed to charge two-part tariff user fees. At the privately optimal interchange fee sellers pay too high merchant fees and buyers pay too low card usage fees compared to the corresponding levels that would be induced by the socially optimal interchange fee.

The intuition for this result comes from the fact that there are two distinct margins, extensive and intensive, on the buyer side and only extensive margin on the seller side. Increasing the IF beyond the socially optimal level not only attracts new cardholders through a higher option value, but also fosters card usage among *existing* cardholders. The incremental buyer surplus due to this extra, inefficient, usage can be extracted at the membership stage through higher fixed fees and lower per-transaction fees, while keeping the average card fee constant. On the seller side, on the other hand, the network cannot fully internalize the incremental seller surplus since sellers make only membership decisions that depend on the average merchant fee.

From the fact that the privately optimal IF is higher than the socially optimal IF, it cannot be concluded that in equilibrium there is an *over-provision* of card services. Improving buyers' usage incentives through a higher IF (inducing, for instance, reward schemes and cash back bonuses) must not necessarily lead to a higher total volume of transactions, since some sellers might abandon the platform in response to higher merchant fees. Hence, from BC's model it can only be concluded that there is *over-usage* in the sense that, in equilibrium, the proportion of buyers who choose to pay by card at an affiliated merchant is inefficiently high.

Note that in the extension of BC's setup where buyers have heterogeneous membership benefits,

[16] Note that the maximization problem of the social planner is the same regardless of the objective being social surplus or user surplus, that is, the maximization of user surplus leads to the maximization of social surplus and *vice versa*. The social planner's maximization problem is analogous to solving for Ramsey fees, like in Rochet and Tirole (2003). Ramsey fees are not driven solely by superelasticity formulae but also reflect each side's contribution to the other side's surplus.

the issuers' fixed fee to buyers is characterized by the Lerner formular. Thus, in line with the result by Rochet and Tirole (2003, 2006) BC's model reveals that a monopolistic issuer introduces a monopoly markup on its fixed costs, inefficiently excluding some buyers from the market.

Result 2 can also be obtained in a weaker form for an extension of BC's benchmark model to the case of a monopoly issuer and a monopoly acquirer. The issuer again sets the card usage fee at its net transaction cost ($f^* = c_I - a$) and the fixed card fee at the option value of cardholding ($F^* = v_B(f)D_B(f)D_S(m)$), thereby extracting exactly the option value of cardholding as a profit. Unlike the previous setting, the acquirer's optimal pricing also involves a markup ($m^* > c_A + a$) which is characterized by the standard inverse elasticity rule over sellers' demand for card services.

As before, a private monopolistic card network sets the IF such that total profits of its member banks are maximized. It still cannot fully internalize the incremental seller surplus since sellers make only membership decisions that depend on the average merchant fee. So the social planner continues to focus more on the sellers' surplus in its objective function and thus puts a higher weight on the acquirer's profit. BC show that the social planner will set a lower IF than the private card network if the pass-through rate of IFs from acquirer to seller is either constant, decreasing in IF or increasing in IF at a sufficiently low rate.

Furthermore, BC show that the result that "Spence distortion" is in favor of buyers because buyers determine the extent of card usage volume, is robust to allowing sellers to surcharge card payments where surcharging is costly. Similarly, the result holds as long as it is too costly for sellers to influence buyers in the form of payment method they use for a given transaction. On the other hand, if surcharging or steering were costless, "Spence distortion" would be corrected. Moreover, "Spence distortion" would be corrected if acquirers were able to apply perfect third degree price discrimination vis-à-vis sellers.

We now compare the second-best (Ramsey) fees, which are induced by the socially optimal IF-maximizing user surplus subject to the banks' optimal fee choice, with the first-best (Lindahl) fees, which the social planner would set if it controlled all user fees. This comparison helps us understand the nature of the externalities in the payment card market and also see whether regulating only IFs would be enough to implement the first-best fees.

The Lindahl fees are derived from the following maximization problem:[17]

$$\max_{F,f,m} W \equiv \{[f + m - c + v_B(f) + v_S(m)]\, D_B(f) D_S(m). \tag{1}$$

BC show that the first-best total price (per transaction) is equal to $c - v_B(f^{FB})$ and is thus lower than the total cost of a transaction. Each type of user is charged a price equal to the cost of a transaction minus a discount, reflecting its positive externality on the other segment of the industry:[18]

$$\begin{aligned} f^{FB} &= c - \left[v_S(m^{FB}) + m^{FB}\right], \\ m^{FB} &= c - \left[v_B(f^{FB}) + f^{FB}\right]. \end{aligned}$$

The socially optimal allocation of the first-best total price per transaction is achieved when the average buyer surplus is equal to the average seller surplus: $v_B(f^{FB}) = v_S(m^{FB})$. Comparing the Ramsey fees with the Lindahl fees BC obtains the following result:

Result 3: By controlling only the IF the social planner cannot implement the first-best (Lindahl) fees.

In practice, the social planner is unable to control all fees in the industry as it can only regulate IFs, which is not enough to achieve full efficiency in the industry. The IF affects only the allocation of the total user price between consumers and sellers, whereas the first-best efficiency also requires a lower total price level due to positive externalities between the two sides.

2.3 Network competition

Rochet and Tirole (2003) show that the impact of introducing network competition (e.g., Visa versus MasterCard) depends on which side of the market users adopt multiple networks (*multi-homing*) rather than a single network (*single-homing*).[19] Intuitively, network competition leads to a bias in the allocation of the total user price favoring the single-homing side since steering users

[17] See Bedre-Defolie and Calvano (2010).

[18] This pricing rule was independently found by Weyl (2009).

[19] This result is also obtained by Guthrie and Wright (2007), Armstrong (2006) and Armstrong and Wright (2007).

toward an exclusive relationship lets platforms extract monopoly rents from the multi-homing side (*competitive bottleneck*).

Casual observation suggests that, at least for the two major card networks, sellers do indeed multi-home. The global card acceptance network of Visa and MasterCard almost perfectly overlaps, with 29 million sellers accepting Visa cards and 28.5 million sellers accepting MasterCards in 2009.[20] Multi-homing is encouraged by a widespread practice called "blending,"which describes the case where acquirers charge one price for accepting different cards from various networks.[21] Besides, Rysman (2007) provides empirical evidence that consumers mostly use only one payment card on a daily basis even when they hold more than one card.

We therefore put an emphasis on describing the effects of network competition when buyers single-home and sellers multi-home. Nevertheless, we also provide a description of the effects of network competition when buyers and sellers multi-home.

□ **Buyers single-home and sellers multi-home**

In the following we describe an extension of BC's benchmark model, which replicates the result by Rochet and Tirole (2003) that network competition leads to a bias in the allocation of the total user price favoring the single-homing side. BC assume two competing three-party card networks (or, alternatively, two competing four-party card networks, in which there is a monopolist issuer and perfectly competitive acquirers). Furthermore, buyers are assumed to be ex-ante heterogeneous in their membership benefits, buyers single-home and sellers multi-home.

For simplicity, BC assume that the networks are homogeneous with respect to the card services they provide, but can be differentiated due to, for example, the brand preferences of users or other differentiated services (such as travel insurance) provided by the card networks. This implies that a buyer receives a membership benefit B_B from holding a card and a convenience benefit b_B from paying by card at a point of sale, regardless of the card platform she uses. Similarly, a seller receives the same convenience benefit b_S from being paid by card regardless of the card network it is affiliated to.[22]

[20] Nilson Report, June 2009.

[21] See the EC's Sector Inquiry, 2007, pp. 16.

[22] As in the benchmark case, BC focus without loss of generality on a per-transaction seller benefit and a per-transaction seller fee.

The timing is modified in the first two stages of the game: First, the platforms simultaneously set their user prices (F_1, f_1) and (F_2, f_2) to buyers and m_1 and m_2 to sellers. Second, buyers realize their membership benefit and decide whether to hold a card and, if so, choose one card network. At the same time, sellers observe their transaction benefit and decide whether to accept the cards of each platform.

Within this setting it can be shown that network competition partially corrects the market power distortion on the buyers' side, but leaves the allocation of the total user price distorted in favor of buyers.

Result 4: If two card networks compete, buyers single-home and sellers multi-home, then the privately optimal price structure comprises a higher merchant fee than the socially optimal price structure. Compared to a monopoly platform, competing platforms charge lower fixed fees to buyers, partially correcting the market power distortion, but retaining the inefficient price allocation between the two user sides.

In equilibrium each network sets the total user price at its cost of a payment transaction, $f_i^* + m_i^* = c$, and allocates the total cost between buyers and sellers to maximize its buyers' card usage surplus since this surplus can be captured by a fixed fee given by the Lerner Formula on the network's residual demand. It follows, when there is network competition, that a network's optimal cost allocation between buyers and sellers coincides with the one maximizing its buyers' card usage surplus and thus with the optimal allocation from buyers' perspective.

In contrast, a social planner would implement a price structure that maximizes the sum of buyers' and sellers' surpluses. For a given total transaction fee, the buyers-optimal merchant fee exceeds the sellers-optimal merchant fee since the average surplus of buyers and the average surplus of sellers are decreasing in their own usage fees. Thus, the price structure chosen by the competing networks leads to a higher merchant fee than would be implemented by a social planner.

Compared to a monopoly platform, competing platforms charge lower fixed fees to buyers because as buyers single-home the competition for buyers increases the price elasticity of each platform's cardholding demand. Platform competition therefore partially corrects the monopoly distortion on the fixed card fee of buyers, but it does not correct the inefficient price allocation between the two sides of users. In fact, Rysman (2007) provides empirical evidence that if buyers

single-home more than sellers, competing payment networks set an IF that is even higher than the IF set by a monopoly network due to the fact that the networks compete more fiercely for buyers. This empirical result suggests that network competition might reinforce the "Spence distortion."

□ **Buyers and sellers multi-home**

Consider now two competing card networks and assume that both buyers and sellers multi-home.[23] Suppose first that as in BC's benchmark case sellers do not internalize the card usage surplus of buyers, and so accept cards only because they get convenience surplus from card acceptance. In that case an increased merchant fee does not affect sellers' demand for the rival network since sellers multi-home. However, a lowered card usage fee decreases buyers' point-of-sale usage of the rival card and thus the rival network's card usage volume. So if a network increases its IF, leading to an increase in merchant fees and a decrease in card usage fees, it exerts a negative externality on the rival's demand. Since competing networks do not internalize this negative externality on the rival's demand, they set higher IFs than a monopoly network. That is, with network competition and buyers and sellers multi-homing the equilibrium IFs are distorted further upward than in BC's benchmark case of a monopoly network.

If sellers do internalize some of the buyers' card usage surplus (for instance, due to the business-stealing effects of accepting cards, like in Rochet and Tirole, 2002; see also section 2.4 on merchant internalization), the negative impact of a lowered card usage fee on the rival network's card usage volume will be partially internalized by sellers, making sellers less willing to accept the rival network's card. This implies that the rival network's card usage volume decreases even further when a network decreases its own IF, exacerbating the negative externality that competing networks exert on each other. Hence, the upward distortion on IFs becomes even more pronounced compared to the case in which sellers do not internalize buyers' card usage surplus.

2.4 Merchant internalization

Another complementary explanation for inefficiently high IFs is merchant internalization. Merchant internalization refers to the fact that merchants can internalize (at least partially) consumer surplus from card transactions, $v_B = E[b_B - f | b_B \geq f]$. This is because by accepting cards they increase

[23] This adoption behavior can be rationalized in equilibrium only if the networks are differentiated.

their service quality, which enables them to increase store demand and/or steal customers from rivals. When merchant internalization holds, merchants accept cards even if the merchant fee is above their transaction benefit: $m > b_S$, which is called as "must-take-cards " by the literature (Rochet and Tirole, 2002, 2011; Bourguignon et al. 2014).

Rochet and Tirole (2002) show that merchant internalization induces a card network to set inefficiently high merchant fees because merchant internalization makes it easier for the network to convince merchants to join them. The greater the merchant internalization, the more likely it is that the card network will exploit the lower "merchant resistance" by setting an inefficiently high IF. Rochet and Tirole (2002) assume that sellers are homogeneous in their transaction benefit from being paid by card. This assumption is equivalent to assuming that third degree price discrimination vis-à-vis (heterogeneous) sellers is possible without friction. Wright (2012) extends their finding to the case of unobserved merchant heterogeneity, which is also assumed by BC and implies that sellers' demand is elastic. Intuitively, merchant internalization makes the merchant demand for card acceptance less elastic to a merchant fee, and so raises the network's optimal IF. The social planner sets a lower IF than the network since it counts consumers' card usage surplus, v_B, only once. Furthermore, Wright (2012) extends this result to network competition.

Result 5: If merchants' internalization of consumers' card usage surplus is sufficiently high, the privately optimal price structure comprises a higher merchant fee than the socially optimal price structure (i.e., the privately optimal IF is higher than the socially optimal one) even if the issuers do not use non-linear card fees.

The necessary assumptions for this result to hold are that the issuer cost pass-through rate is not much higher than the acquirer cost pass-through rate and merchants are not allowed to surcharge card payments against cash (see Wright, 2012).

3 Regulation of interchange fees

The theoretical arguments summarized above suggest that IFs help internalize the complementarity between services on both sides of the payment card market, but private platforms will set higher IFs than is socially optimal. In line with that, policy makers are concerned that IFs restrict

competition between acquiring banks and inflate the costs of card acceptance by sellers without improving efficiencies. IFs are seen as a form of price fixing among horizontal competitors (see e.g., EU MasterCard decision, 2007, para. 405). Sellers may be exploited ex post as they may accept a payment card, even if it is more expensive than other payment instruments, in order to increase service quality. They would accept cards as long as the merchant fee is lower or equal to their own net benefits plus their buyers' net benefits (in line with the merchant internalization argument, see section 2.4).

With regard to efficiencies, the EU Commission acknowledged in its MasterCard decision that payment systems can be characterized by indirect network externalities and that, in theory, IFs can help optimize the utility of a card network to all of its users. However, in practice MasterCard could set its IFs using largely arbitrary and inflated cost benchmarks. Further, MasterCard has failed to submit empirical evidence that by pursuing its member banks' aim of maximizing sales volumes its IFs have created efficiencies that benefit all customers, including sellers. What is more, the EU Commission raised as an argument that the existence of payment schemes that function without IFs would demonstrate that IFs are not indispensable for the viability of payment cards. In fact, ECB statistics would indicate that card schemes without IFs display the highest card usage per capita in the EU.[24]

These concerns have led many authorities to cap IFs, for example, in Australia, Canada, Chile, Denmark, Mexico, Singapore, Switzerland, and the US. Other policy makers, among them the EU Commission, plan to cap IFs or have enforced caps through commitment decisions.[25] However, determining caps on IFs is non-trivial in practice, even when using simple models, *inter alia* because an efficient fee structure does not necessarily reflect the relative costs of the two sides of the market. The average surplus and elasticities on both sides need to be considered, but they are hardly observable and measurable.

When determining caps on IFs, policy makers often apply a test proposed by Rochet and Tirole (2011) referred to as the "Merchant Indifference Test," "Tourist Test," or "Avoided Cost Test." Under this test IFs that make sellers indifferent between a transaction by card and a transaction

[24]Citing EU Commission (2007b); see also Brestam and Schmiedel (2011), who show differences in card payment rates in Annex 1 as well as the number of card payment transactions per capita in the euro area in chart 3.

[25]On 20 April 2015 the Council of the EU adopted a regulation capping interchange fees for payments made with cards. With certain exceptions, the maximum levels are 0.3% of the value of the transaction for all credit card transactions and 0.2% of the value of the transaction for all debit card transactions.

in cash are assessed to be efficient. These IFs correspond to the value of the cost savings that card use generates for sellers in comparison to cash payments. Rochet and Tirole argue that to the extent that certain conditions are met, IFs that pass the test ensure that cardholders make efficient choices with respect to payment instruments.

In particular, Rochet and Tirole (2011) show that the test is an exact test of excessive IFs from the point of view of total user surplus when issuers' margin is constant. Note that issuers' margin is constant, for example, when issuers are perfectly competitive but also in other cases of issuer market power. However, when issuers' margin is variable, the IF that maximizes total user surplus is higher than the IF that meets the test in the cost amplification case and lower than the IF that meets the test in the cost absorption case.[26] Further, the test is no longer an exact test of excessive IFs in case of distortions. Rochet and Tirole (2011) provide as an example of a distortion resulting from unobserved seller heterogeneity as assumed in BC's model. This distortion cannot be corrected for through bargaining between sellers and buyers as the card fee does not depend on the sellers' identity.

With regard to the necessity of capping IFs, Rochet and Tirole point out that IFs that pass the test only coincide with the privately optimal IFs when two networks are competing and all consumers multi-home. In all other cases, for example, when some consumers hold single cards (or hold multiple cards but use only a single card), when there is a monopoly platform, where issuers exert some market power or in case of distortions the privately optimal IFs are set higher than the IFs that pass the test, and may therefore be set too high from the point of view of users, making policy intervention necessary.[27]

To conclude, many policy makers have shown concern as it is likely that the privately optimal IFs are set too high compared to the socially optimal level, making policy intervention necessary. The test of excessive IFs, proposed by Rochet and Tirole (2011) and applied by many policy makers, is appealing because of its simplicity but it might be an imperfect measure to identify excessive IFs in the various environments summarized above.

[26] Cost absorption means the issuers' margin decreases with costs. Cost amplification means the issuers' margin increases with costs.

[27] As explained in section 2.3, even when two networks compete and all consumers multi-home, IFs are too high from the point of view of users in case of sellers differing in their net benefits derived from card transaction (described as distortion by Rochet and Tirole, 2011, and assumed by BC).

4 The impact of the no-surcharge-rule on social welfare

Most theoretical models on the pricing of payment cards assume that sellers do not surcharge card payments. This assumption seems to hold in practice as either card networks impose the no-surcharge rule (NSR) or sellers hardly make use of the possibility to surcharge.[28] If surcharges of card payments are observed at all, then it is for purchases on the Internet such as for card payments of flight tickets. In these cases, surcharges often exceed the actual merchant fees, suggesting that sellers only use surcharging to implement a form of add-on pricing.

When assessing the welfare effect of the NSR it should be considered that the NSR is a price restriction imposed by the card payment network on its sellers. The card payment network supplies an infrastructure to sellers and with the NSR it conditions sellers' price for the payment method on the prices of rival payment methods. Hence, the NSR can be seen as a vertical restraint in a two-sided market. In a vertical relationship, if the supplier restricts the pricing of its products by the retailer (e.g., by a Resale Price Maintenance), this will raise anti-competitive concerns. It is, however, unclear how vertical restraints should be addressed in two-sided markets.

Theoretically, the inherent asymmetry in decisions between cardholders and sellers might be equalized when surcharging is possible as surcharging enables sellers to pass-through merchant fees on to buyers, which in turn make buyers internalize sellers' card acceptance costs. If that is the case, by setting a higher IF the card network will no longer be able to induce higher card usage because a higher IF will inevitably lead to higher surcharges. Rochet and Tirole (2002) show that if surcharging is possible without friction, the level of IF becomes neutral. To understand this neutrality result, consider first a three-party card network. Sellers who accept card payments only care about the net cost of accepting card payments (i.e., the merchant fee minus the surcharge). Similarily, cardholders only care about the total cost of paying by card (i.e., the per-transaction fee plus the surcharge). It does not matter how the three-party card network distributes its fees between buyers and sellers since sellers can reflect their merchant fee in their surcharge. Gans and King (2003) illustrate that the neutrality result indeed holds for more general frameworks of four-party card networks regardless of the degree and type of issuer competition or acquirer

[28] Several countries decided to prohibit the NSR. In the US, it was only recently found that the NSR violates antitrust law. In April 2015, a New York federal judge decided that AmEx must change its anti-steering rules that barred merchants from encouraging consumers to use cheaper forms of payments. He also ruled that merchants could offer discounts and other incentives to customers for using cards with lower fees.

competition or merchant competition: if a network increases its interchange fee it will increase its merchant fee and decrease its cardholder fee. As long as sellers can set different prices to different payment methods, they will surcharge card payments so that the total cost to sellers of accepting card payments and the total cost to buyers of paying by card do not depend on how the total card transaction fee is allocated between sellers and buyers, that is, the level of IF will have a neutral effect.

It is puzzling why surcharging is seldomly observed in practice. One reason could be that surcharging is costly to sellers. Surcharging might lead to administrative costs but might also induce fairness concerns. Administrative costs might occur as, depending on the chosen payment method, an additional surcharge must be added to the total payment amount at the checkout. The buyer must be informed of the surcharge and reconsider its decision on the payment method. If the optimal surcharge from sellers' perspective was rather small, the cost of surcharging could exceed the sellers' benefit from surcharging. What is more, sellers' decision to not make use of surcharging might be driven by fairness concerns. If the optimal surcharge from sellers' perspective was rather small, sellers might want to avoid surcharging so as to appear more customer- and service-oriented. This might also be particularly valid if competitors do not surcharge, buyers are unaware of the additional fees that sellers incur when being paid by card instead of cash, and buyers are aware of sellers' surcharging policy before visiting the store.

The theoretical literature documents that the effect of the NSR on social welfare is ambiguous. Most of the literature assumes that buyers' consumption demand is inelastic (i.e., buyers always consume one unit of the good) and focus on how allowing or banning NSR affects buyers' card usage decision. Wright (2003) finds that if sellers are perfectly competitive and homogeneous, the NSR has no impact on the transaction volume or social welfare. If, on the other hand, sellers are monopolistic and homogeneous, the NSR increases the volume of card transactions and social welfare since it prevents sellers' ex-post monopoly mark-up limiting card usage. Considering a setting in which sellers are homogeneous (no unobserved heterogeneity) and imperfectly competitive (Hotelling-Lerner-Salop model) Rochet and Tirole (2002) show that the impact of the NSR on social welfare is ambiguous. With large issuer market power the NSR is more likely to be welfare increasing. This is because without the NSR, there is under-provision of card services due to issuer market power. With the NSR, there is over-provision of card services since the inefficiently high IF induces a too

low card fee. Issuer market power then increases merchants' resistance as it leads to an increase of the card fee. Issuer market power therefore decreases the extent of inefficiency due to too high IFs when the NSR applies. Wright (2012) points out that if surcharging is costly and sellers differ in their costs, the result of merchant internalization leading to an increase in IFs holds. He also finds that in this setup when merchants are homogeneous in their transaction benefits, issuer and acquirer margins are constant and very close to zero, the network would set a higher IF than the socially optimal one. Very recently, Edelman and Wright (2014) show that a monopoly intermediary always prefers to impose price coherence (uniform price regardless of purchasing channel) on its sellers, this reduces the consumer surplus and sometimes the total welfare due to an over-consumption of the intermediary's service and also due to an over-investment of intermediary in buyer-side benefits. They also find that competition among intermediaries intensifies these distortions.

Very few papers look at the situation where buyers' consumption demand is elastic and so consumption of the good might also be affected by the card fees and whether NSR is allowed or banned. Schwartz and Vincent (2006) show that NSR increases card transactions and reduces cash transactions and so NSR increases the total welfare if and only if there is a sufficiently large amount of cash users. They assume exogenous amount of card users and cash users and so they cannot analyze how NSR affects consumers' choice of the means of payment, but they instead analyze how NSR impacts the consumption decisions of cash users and cardholders. Bourguignon, Gomes, and Tirole (2014) model merchant surcharging policy as an add-on that consumers are not informed about before visiting the store (like in Ellison, 2005) and so surcharging creates a hold-up problem when visiting a store is costly. In this setup merchants resist card acceptance less since they do not want to miss sales at a point-of-sale (ex-post merchant internalization). Assuming homogeneous merchants they show that banning surcharging increases welfare if the merchant fee is sufficiently high (above the tourist test level) and decreases welfare otherwise. When surcharging is allowed, capping merchant fees is welfare reducing. This suggests that the optimal policy toward the NSR is related to public policy toward merchant fees of IFs.

Schuh, Shy, and Stavins (2010) point out that the NSR can have a distributional effect, which might not be desired. Under the NSR, sellers can only pass on the costs of card acceptance to buyers by setting higher retail prices which are paid by card users and cash users. Hence, the NSR leads to redistribution from ("less wealthy") cash users to ("more wealthy") card users.

5 Concluding remarks

In this overview article we have summarized the main sources of market failure due to payment networks' pricing that the theoretical literature has identified. Given that sellers are mostly not allowed to or do not price discriminate based on the type of the payment method, the payment card usage volume, profits, and user surpluses depend on the allocation of total user fees between sellers and buyers (the price structure), like in other two-sided markets and different from a standard one-sided market. In a three-party (closed) card network the network (like AMEX, Discovery) determines all prices, so the total user price as well as the allocation of this price between sellers and buyers. However, in a four-party (open) card network (like Visa, MasterCard) the network can control the price structure indirectly through an interchange fee paid by the merchant's bank to the acquirer's bank and banks set prices to final users. One common finding is that in a four-party card scheme an interchange fee can help internalize the complementarity between services on both sides of the market. However, the literature identifies two types of distortions due to the pricing of privately optimal platforms. Firstly, the market power of a platform leads to higher total user prices similar to the case of one-sided markets. Secondly, the private platform's choice of the price structure (or interchange fee) differs from the one of the social planner. Moreover, it is shown that platform competition can help to correct for such market power distortion, but may exacerbate the distortions with regard to the allocation of the total user price.

Although the early literature could not drive clear conclusions on the direction of the price structure distortion, the recent theoretical literature identifies three complementary sources of market failure in this industry due to the private platform setting too high merchant fees and too low card user fees (or too high interchange fees) compared to the social planner: 1) Asymmetric decisions of buyers and sellers, 2) merchant internalization, 3) platform competition. Too high merchant fees (or interchange fees) may even result in the case of very little market power of the issuing or acquiring banks.

When sellers cannot surcharge card payments without any friction, cardholders are the ones who determine the extent of card usage at a merchant location accepting cards. In other words, buyers decide on membership and usage, whereas sellers decide only on membership. This asymmetry between the decisions of buyers and sellers imply that the banks can (fully) internalize

buyers' card usage surplus, but not sellers' card usage surplus, by setting two-part tariffs. As a result, the private platform favors buyers too much at the expense of sellers: over-taxing sellers and under-pricing buyers (via a too high interchange fee in a four-party card scheme).

Merchant internalization means that merchants may accept cost-increasing cards as a way to increase their store demand and/or steal customers from their rivals and/or not to lose consumption at a point-of-sale. The greater the merchant internalization, the easier it is to convince merchants to join the network, and the more likely it is that card networks will exploit the lower merchant "resistance" to fee increases by setting inefficiently high merchant fees.

A platform's pricing structure will be distorted in favor of buyers if two networks compete and buyers single-home more than sellers, because then the networks will try to compete for buyers by lowering their fees and will charge merchants the monopoly price for providing an exclusive access to their cardholders. If both buyers as well as sellers multi-home, the privately optimal price structure will induce a lower price for buyers and a higher price for sellers (or a higher interchange fee) than in case of a monopoly network.

The theoretical literature further reveals that a regulator will be unable to implement the first-best (Lindahl) fees, as it cannot control for all fees in the market. Regulating the interchange fee is shown to not be enough to achieve full efficiency in the industry. The interchange fee affects only the allocation of the total user price between buyers and sellers whereas the first-best efficiency also requires a total price level lower than a marginal cost of a transaction due to positive externalities between the two sides. In principal, regulating the interchange fee might enable the planner to implement the second-best (Ramsey) fees, but this has been proven to be very difficult in practice as they depend on factors that are very difficult to measure, such as the demand elasticities of the two user sides, the average benefits of buyers and sellers from card payments, and issuer and acquirer cost pass-through rates. Even if the literature identifies the three complementary reasons that might lead to too high interchange fees (as we discussed above), the literature provides no basis for a cost-based cap regulation on interchange fees.

A simple test for excessive interchange fees was proposed by Rochet and Tirole (2011) and applied by many policy makers thereafter. An interchange fee (which induces a certain merchant fee) passes the test if and only if accepting a card payment does not increase the sellers' operating cost. However, the proposed test is only an exact test of excessive interchange fees from the point of

view of total user surplus when issuers' margin is constant and when sellers do not differ in their net benefits derived from card transactions (or when there is no unobserved merchant heterogeneity).

Most theoretical models on the pricing of payment cards assume that sellers cannot or do not surcharge card payments. This assumption seems to hold in practice as either card networks impose the no-surcharge rule (NSR) or sellers seldomly surcharge card payments in practice even when the NSR is not imposed. It needs to be better understood why sellers seldomly surcharge card payments in practice even if they do not face NSR. From a theoretical point of view the effect of the NSR on social welfare is ambiguous. One puzzling question is why many merchants do not surcharge even when they are allowed to do. In particular, more empirical research needs to be done to understand the determinants of merchants' decisions of whether to surcharge.

There are many important questions left for future research and policy discussions. One important question is how banks would react to a card fee regulation and what the resulting effect would be on consumer and merchant welfare. Bedre-Defolie, Song and Ullrich (2015) have started to analyze the short-run reaction of banks using the national debit card scheme data in Norway. This is the first analysis of estimating demands for payment cards using prices and market shares of differentiated banks to consumers and merchants. There might also be long-run reaction by changing investment in infrastructure, quality (Verdier, 2010). If the price structure implies a too high interchange fee for open networks, it should also imply too high merchant fees for closed networks. It is unclear how a cap regulation on interchange fees affects the competition between open networks that are subject to the regulation and closed card networks which do not have explicit interchange fees. The impact of an interchange fee on acquirer competition or on issuer competition needs to be further analyzed. For instance, Bedre-Defolie and Calvano's (2013) extension of imperfect issuer competition illustrates how an interchange fee could be used strategically to raise fixed (annual) card fees for consumers, and so soften issuer competition.

References

Armstrong, M. (2006): "Competition in Two-Sided Markets," *The RAND Journal of Economics*, 37(3), 668–681.

Armstrong, M., and J. Wright (2007): "Two-Sided Markets, Competitive Bottlenecks and Exclusive Contracts," *Economic Theory*, 32(2), 353–380.

Baxter, W. (1983): "Bank Interchange of Transactional Paper: Legal and Economic Perspectives," *Journal of Law and Economics*, 26(3), 541–588.

Bedre-Defolie, O., and E. Calvano (2010): "Pricing Payment Cards," ESMT Working Paper No. 10-005 (R2).

——— (2013): "Pricing Payment Cards," *American Economic Journal: Microeconomics*, 5(3), 206–231.

Bedre-Defolie, O., M. Song, and H. Ullrich (2015): "Assessing the Impact of Payment Card Fee Regulation," Working Paper.

Börestam, A., and H. Schmiedel (2011): "Interchange Fees in Card Payments," *European Central Bank Occasional Paper Series No. 131.*

Bourguignon, H., R. D. Gomes, and J. Tirole (forthcoming): "Shrouded Transaction Costs," *Quarterly Journal of Economics.*

Brenning-Louko, M., T. Panova, L. Repa, and A. C. Teixeira (2006): "Interchange Fees and Incentives to Invest in Payment Card Systems," *EC Competition Policy Newsletter*, 2, 12–14.

Caillaud, B., and B. Jullien (2003): "Chicken & Egg: Competition Among Intermediation Service Providers," *The RAND Journal of Economics*, 34(2), 309–328.

Chakravorti, S. (2010): "Externalities in Payment Card Networks: Theory and Evidence," *Review of Network Economics*, 9(2).

Edelman, B. G., and J. Wright (forthcoming): "Price Coherence and Excessive Intermediation," *Quarterly Journal of Economics.*

Ellison, G. (2005): "A Model of Add-On Pricing," *Quarterly Journal of Economics*, 120, 585–637.

European Commission (2007a): "Final Report on the Retail Banking Sector," Working Paper.

——— (2007b): "Antitrust: Commission Prohibits MasterCard's Intra-EEA Multilateral Interchange Fees," MEMO/07/590.

Evans, D., and R. Schmalensee (2005): "The Economics of Interchange Fees and Their Regulation: An Overview," Conference Interchange Fees in Credit and Debit Card Industries: What Role for Public Authorities.

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

Guthrie, G., and J. Wright (2007): "Competing Payment Schemes," *The Journal of Industrial Economics*, 55(1), 37–67.

Hayashi, F. (2009): "Do U.S. Consumers Really Benefit from Payment Card Rewards?," *Economic Review*, Q I, 37–63.

Laffont, J., S. Marcus, P. Rey, and J. Tirole (2003): "Internet Interconnection and the Off-Net-Cost Pricing Principle," *The RAND Journal of Economics*, 34(2), 370–390.

Rochet, J.-C., and J. Tirole (2002): "Cooperation among Competitors: Some Economics of Payment Card Associations," *The RAND Journal of Economics*, 33(4), 549–570.

——— (2003): "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, 1(4), 990–1029.

——— (2006): "Two-Sided Markets: A Progress Report," *The RAND Journal of Economics*, 37(3), 645–667.

——— (2011): "Must-Take Cards: Merchant Discounts and Avoided Costs," *Journal of the European Economic Association*, 9(3), 462–495.

Rysman, M. (2007): "An Empirical Analysis of Payment Card Usage," *The Journal of Industrial Economics*, 55(1), 1–36.

Schmalensee, R. (2002): "Payment Systems and Interchange Fees," *The Journal of Industrial Economics*, 50(2), 103–122.

Schuh, S. D., O. Shy, and J. Stavins (2010): "Who Gains and Who Loses from Credit Card Payments? Theory and Calibrations," Public policy Discussion Papers, Federal Reserve Bank of Boston.

Schwartz, M., and D. R. Vincent (2006): "The No Surcharge Rule and Card User Rebates: Vertical Control by a Payment Network," *Review of Network Economics*, 5(1).

Shy, O., and Z. Wang (2011): "Why Do Payment Card Networks Charge Proportional Fees?," *The American Economic Review*, 101(4), 1575–90.

Verdier, M. (2010): "Interchange Fees and Incentives to Invest in Payment Card Systems," *International Journal of Industrial Organization*, 28(5), 539–554.

——— (2011): "Interchange Fees in Payment Card Systems: A Survey of the Literature," *Journal of Economic Surveys*, 25(2), 273–297.

Wang, Z. (2010): "Market Structure and Payment Card Pricing: What Drives the Interchange?," *International Journal of Industrial Organization*, 28(1), 86–98.

Weyl, E. (2009): "Monopoly, Ramsey and Lindahl in Rochet and Tirole (2003)," *Economics Letters*, 103(2), 99–100.

——— (2010): "A Price Theory of Multi-Sided Platforms," *The American Economic Review*, 100(4), 1642–1672.

Weyl, G., and A. White (2010): "Imperfect Platform Competition: A General Framework," Mimeo, Harvard Economics.

Wright, J. (2003): "Optimal Card Payment Systems," *European Economic Review*, 47(4), 587–612.

——— (2004): "The Determinants of Optimal Interchange Fees in Payment Systems," *The Journal of Industrial Economics*, 52(1), 1–26.

——— (2012): "Why Payment Cards Fees Are Biased Against Retailers," *The RAND Journal of Economics*, 43(4), 761–780.