TRANSACTION COSTS, EXTERNALITIES, AND "TWO-SIDED" PAYMENT MARKETS

Dennis W. Carlton*

Alan S. Frankel**

I. INTRODUCTION

Recent study of two-sided markets has generated a rapidly growing and technically complicated body of

* Professor Carlton teaches at the University of Chicago Graduate School of Business.

** Mr. Frankel is a Senior Vice President at Lexecon, an economics consulting firm.

The authors wish to thank Gustavo Bamberger, Kevin Murphy, Allan Shampine, Jean Tirole, and participants of the April 2005 conference on two-sided markets held at the Columbia University School of Law for their helpful discussions. Both Messrs. Carlton and Frankel have served as experts on behalf of clients adverse to Visa and MasterCard.

literature. As we will describe more precisely in this article, a two-sided market is one in which externalities exist amongst market participants. Domestic and international antitrust litigation and regulation involving payment system networks, such as those operated by the MasterCard and Visa bank joint ventures, have contributed to the increased interest in two-sided markets. Central to the issues being litigated are the collective establishment of policies, such as "no-surcharge rules," that govern how retail merchants may set prices and the setting of transfer payments ("interchange fees") between the members of the payment system networks.

In this paper, we review the fundamental economic forces that underlie two-sided markets and then discuss how these forces relate to current debates over the competitive effects of certain collective policies, such as the setting of interchange fees. Visa, MasterCard, and some economists have put forward pro-competitive justifications for these policies based on the "two-sidedness" of payment systems.[1] We pose and discuss a series of questions that probe the empirical and policy relevance of these justifications.

II. TRANSACTION COSTS AND EXTERNALITIES

Transaction costs and externalities represent the two economic forces at the heart of two-sided markets analysis. Because externalities exist only in the presence of transaction costs, these concepts are invariably interrelated.

A. Transaction Costs

Transaction costs are the expenses necessary to effect the transfer of goods from a seller to a buyer. The transaction costs that result from the purchase, installation, and usage

[1] *See, e.g.*, William F. Baxter, *Bank Interchange of Transactional Paper: Legal and Economic Perspectives*, 26 J.L. & ECON. 541 (1983); Jean-Charles Rochet & Jean Tirole, *Two-Sided Markets: An Overview* 1 (Indus. Econ. Inst., Working Paper No. 275, 2004), *available at* http://www.idei.fr/doc/by/tirole/rochet_tirole.pdf.

of scales and weights, cash registers, and electronic point-of-sale terminals influence consumer welfare in much the same way as the costs associated with building a factory. Consumer welfare is clearly affected by changes in these costs, but merely identifying such costs as "transaction costs" does not necessarily require a fundamental rethinking of the nature of competition and pricing.

Assuming that total transaction costs stay constant and selling prices can (and do) adjust, buyers and sellers are indifferent as to which party must bear the transaction cost. For example, if it costs $1 to pay a clerk to process Bob's $9 transaction at Wal-Mart, it does not matter whether Bob pays $1 directly to the clerk or instead pays $10 to Wal-Mart, and Wal-Mart pays the clerk $1. Bob's direct cash outlay and Wal-Mart's net cash inflow remain the same. So long as the prices to individual consumers stay flexible, the allocation of transaction charges between the parties in this example remains "neutral."

The cost allocation should be theoretically neutral, but sometimes the economic reality is quite different. This may occur because one side has a comparative advantage in dealing with transaction costs.[2] Collection of sales tax proceeds represents a simple example. Instead of requiring the retail merchant to remit tax revenue, the government could instruct individual consumers to send in payments in connection with their retail spending. Although these two alternatives would presumably result in equivalent tax revenues, in practice the collection and compliance costs will likely be greater under the latter option. Therefore, it is more efficient for the retail merchant, rather than the consumer, to remit the tax proceeds. Similarly, in the Wal-Mart example, Wal-Mart should pay the checkout clerk because the store is in a better position to monitor worker efficiency.

[2] Information costs are related to transaction costs. Holding constant the amount of the charge and the party on whom it is assessed, the form of a charge may influence its effect in the market. For example, a $1 fee charged to a customer at the end of a month may be less noticeable than a $1 fee imposed at the point of sale.

If transaction costs differ across consumers, such that it is more costly to transact with some consumers than others, then three main outcomes are possible: (1) sellers can charge consumers differently according to the transaction costs a consumer imposes; (2) firms can specialize in serving a particular type of consumer group and price accordingly; or (3) a pooling equilibrium may develop, in which the seller serves all consumers but charges a single price. Pooling results, in part, from economies of scale and transaction costs. When many different consumer types exist, having different firms serve each type of consumer may waste resources. As the market grows, however, specialization often increases.[3] When the transaction costs of price discrimination are high, firms may charge a single price to all consumers. Although select consumer groups may impose different costs on a firm, a business may choose to charge all customers equally to avoid the transaction costs of establishing separate pricing schemes.

Even in what economists view as highly competitive markets, transaction costs limit the flexibility of the pricing system to signal scarcity. Supermarkets, for example, do not vary prices continuously according to inventories on hand or the length of queues at checkout counters. Of course, technological advances have expanded the potential pricing practices available to businesses. To illustrate, some states provide lower priced toll lanes for transponder-equipped vehicles.[4] The toll authorities could use computer monitoring sensors embedded in the pavement to monitor traffic speed in conjunction with signs to announce the current price for driving on a toll road. Establishing fixed prices helps consumers predict and evaluate their consumption decisions. In the highway example, however, congestion-based pricing (or time-of-day pricing) can provide an additional incentive

[3] George J. Stigler, *The Division of Labor is Limited by the Extent of the Market*, 59 J. POL. ECON. 185, 188 (1951).

[4] For example, in Illinois cash tolls are twice those paid by transponder-equipped vehicles. *See* Illinois State Toll Highway Authority, http://www.illinoistollway.com.

for motorists to alter their travel schedules or routes so as to decrease the costs imposed on other motorists. A similar idea lies behind time-of-day pricing for electricity consumption.

Refining prices is impractical if the costs of such exercise exceed the savings from more accurate price signals. If it costs less, for example, to build and use additional peak load electrical generators on hot summer days than to install and operate the equipment necessary to charge households according to time and temperature, then refined pricing may be inefficient. Instead, a pooling equilibrium will exist in which different consumers (e.g., peak and off-peak) are treated equally. In general, economic policy should not interfere with the reduction of transaction costs that impede the realization of pricing efficiencies.

New transaction technologies can alter the pricing equilibrium that exists in a given market. For example, until the late 1990s, most passengers purchased airline tickets through independent travel agents.[5] Pooling equilibrium existed in which consumers paid the same price for a ticket irrespective of how intensely they used the travel agent's services. Airlines paid travel agents according to each ticket sold. Although the level of service that customers received varied, it was more efficient for airlines to adopt a uniform commission structure that disregarded the amount of time agents spent serving particular customers.[6] In this case, the total cost to consumers was no different if they paid a ticket price that included the travel agent's commission than if they paid a separate commission to the travel agent (and a correspondingly lower ticket price).

The proliferation of airline and travel websites has significantly reduced the costs of searching for and booking flights for Internet-savvy consumers. Moreover, airlines can

[5] Severin Borenstein & Nancy L. Rose, Nat'l Bureau of Econ. Research Conference on Regulation, *Regulatory Reform in the Airline Industry* 13 (2005), *available at* http://www.nber.org/~confer/2005/Reg/rose.pdf.

[6] It would have been difficult for airlines to monitor and to compensate independent travel agents in amounts that varied according to the level of service provided to individual travelers.

directly monitor which consumers utilize the Internet to purchase tickets. Because using the Internet significantly lowers transaction costs, airlines have reduced the incidence of price pooling by offering lower net prices to Internet customers and by abolishing commissions paid to travel agents.[7] Consumers can still receive personalized services from ticket agents, but they typically must pay an additional service charge directly to the agent. Accordingly, consumers employ travel agents only if the value of the agent's services exceeds the service fee. In this way, price refining efficiently steers consumers towards the lower cost alternative in the same way that self-serve and full-serve gasoline pump pricing did twenty years earlier. Thus, the Internet has caused a movement away from a pooling equilibrium to a specialized equilibrium.

B. Externalities

Externalities arise when the private cost facing a buyer or seller differs significantly from the social cost. Factory emissions and highway congestion illustrate classic examples. If a factory owner does not have to compensate his or her neighbors for the pollution costs the factory imposes upon them (e.g., for cleaning, health care, and a lower quality of life), then the factory has an economic incentive to overpollute. Similarly, a motorist who does not have to compensate other motorists for slowing traffic has an incentive to drive too much.[8] By taxing pollution or charging highway tolls in amounts equal to the social costs imposed by these activities, regulators can theoretically correct these market failures.

Whereas the traditional "Pigouvian" view of externalities describes the polluting factory as unambiguously imposing costs on others, Ronald Coase challenged this notion by

[7] Borenstein & Rose, *supra* note 5, at 13-14.

[8] Congestion externalities, such as driving too often on a highway, are examples of negative "network externalities." *See infra* Section II.C for further discussion.

acknowledging the symmetric nature of externalities.[9] While a factory may impose costs on nearby homeowners, the homeowners create such costs by choosing to reside near sources of pollution. It might be cheaper for society to relocate the homeowners or to compensate them for accepting dirtier air than to force the factory to move or reduce pollution. Having the authority to regulate air quality would seem to correspond to the ability to reduce pollution, but Coase argued that this assumption might be flawed in certain circumstances.[10]

Coase explained that externalities and transaction costs are related phenomena.[11] In a world with no transaction costs, an efficient allocation of resources will occur irrespective of the initial allocation of property rights and the existence of the externality. The allocation of costs, rights, and obligations is "neutral" in such a hypothetical world, and private contracting is sufficient to solve the externality problem. Coase's insight, though profound, is simplistic: regardless of the allocation of property rights, the efficient solution will produce more output that can be used to make each party better off (or at least one party better off without harming the other).[12] For this reason, the most efficient solution will always be attained absent transaction costs.

Of course, it may not be possible to overcome transaction costs. If there are many parties to a potential negotiation, for example, high bargaining costs might impede private resolution. Given the existence of transaction costs, the initial assignment of property rights affects the allocation of resources. Market failures occur when parties cannot resolve externalities privately due to transaction costs. In some cases, the potential for market failure makes market regulation or an initial allocation of property rights sensible.

[9] Ronald Coase, *The Problem of Social Cost,* 3 J.L. & ECON. 1 (1960).

[10] *Id.* at 17-19.

[11] *Id.* at 2-8.

[12] *Id.* at 5-6.

One solution might be to assign financial responsibility for external costs to the party best able to control them.

Some externalities arise in the context of commercial relationships. For example, discounters often free-ride off of the promotional efforts and services that full-service retailers provide. If sellers could charge customers directly for their services and their promotional efforts, whether or not customers purchase the good, then there would be no externality. If this pricing structure is not possible, then a distributor may choose not to promote a manufacturer's products in the presence of free-riding because the distributor bears all of the costs but does not reap all of the benefits from his or her promotional efforts. In such a case, the distributor may use vertical contractual restrictions, such as exclusive territories, to reduce free-riding. Exclusive territory provisions allow a promoting distributor to sell to customers who received promotional services. An exclusive territory presents a tradeoff to a manufacturer: it reduces the inefficiencies of free-riding among distributors, but also reduces the competition among them. In addition, some potential customers of a discounter may be forced to pay higher prices even though they may not need costly selling services.

C. Network Externalities

In a typical market that is not characterized by externalities, an individual's demand can be described by:

$$Q_i = f(P, Z).$$

The quantity, Q_i, that an individual demands depends on the price, P, and other factors, Z, such as income and other prices (we henceforth suppress Z). A direct "network externality" exists when the consumption benefit to one consumer depends not only on the price, but also directly on the number of other consumers who consume the good. Hence,

$Q_i = f(P,Q)$, where Q is total consumption.

Common examples of markets that are characterized by network externalities include telephone services, fax machines, and email. For example, an individual's demand for telephone service depends on how many others have a phone. The presence of a network externality in a competitive industry might suggest that a subsidy or added incentive is appropriate. Early adopters perhaps should be subsidized because their decision to participate could significantly increase the value of participating to others. No one supplier can take into account the snowball effect that early adopters might have on the growth of the industry. In the extreme case, there may be a "chicken and egg" problem in which no one makes the initial investments necessary to induce everyone else to join the network, and the technology remains unexploited.

Many products that arguably exhibit network effects, however, are adopted successfully without any subsidy.[13] Often, some firms choose to invest more heavily than others in the early stages of development because they hope to achieve widespread product adoption and to realize a significant share in the resulting market. Users can often internalize some of the benefits others receive; a business installing a fax machine, for example, will benefit as its costs of conducting business decline. It is uncertain how many industries have significant network externalities that persist for long periods of time.[14] Many network products or services are adopted quite rapidly by large segments of consumers.[15] Still other products are adopted after a more prolonged startup phase, and some product markets might never develop at all.

[13] S.J. Liebowitz & Stephen E. Margolis, *Network Externality: An Uncommon Tragedy*, 8 J. ECON. PERSP. 133, 137-38 (1994).

[14] *See, e.g.*, Allan Shampine, *Compensating for Information Externalities in Technology Diffusion Models*, 80 AM. J. AGRIC. ECON. 337 (1998) (showing limited opportunity for benefits from a social planner's attempt to compensate for information externalities).

[15] For example, consumers adopted fax machines so rapidly that some early business ventures that sought to offer fax services to those that lacked their own machines quickly failed.

D. Indirect Network Effects and Two-Sided Markets

An indirect network effect exists when one member of a group ("Group A") does not *directly* value the other members of Group A, but values them *indirectly* because Group A member presence induces additional participation by members of a highly valued "Group B." For example, in the context of payment systems, consumers benefit when more merchants accept the credit card brand that they carry, and merchants benefit when more consumers carry the card that they accept.

As stated, however, this indirect "network effect" may be true of markets throughout the economy. For example, a consumer of fresh salmon benefits indirectly from the presence of other salmon eaters. By collectively providing a larger incentive for fishermen to bring salmon to fish markets, salmon consumers reap the gains from economies of scale. More generally, consumers often benefit when more merchants carry the products they like, and merchants often benefit when more consumers seek the products they sell. Liebowitz and Margolis have cautioned that this apparent indirect network effect can be an example of a *pecuniary* externality—i.e., one handled efficiently by decentralized, competitive pricing, rather than one that results in market failure, and correspondingly justifies market intervention.[16] It is not an externality, for example, when one consumer's purchases drive up the price that a second consumer pays for a given product because the resulting price increase reflects the true cost and value of the units at the margin.

When Product A has a complement Product B in consumption, demand for A depends not only on the price of A, but also on the price of B, which may be a function of the output of B. If there are scale economies in the production of B, then:

$$Q_A = f\,(P_A, P_B(Q_B)).$$

[16] Liebowitz & Margolis, *supra* note 13, at 136-39.

As explained, products with complements or substitutes in consumption generally give rise to pecuniary effects only, and ultimately the competitive equilibrium is efficient. Pricing efficiency does not demand that a subsidy be given to jelly producers so that the benefit of lower jelly prices is transferred to peanut butter consumers.

Suppose the demand for A directly depends not only on the price of A, but also on the quantity of B:

$Q_A = f(P_A, Q_B)$ (1).

As an illustration, the demand for credit cards may depend both on their price and the number of merchants that accept them. The merchants' demand for card acceptance services may depend on the price that they are charged, and the number of consumers who carry the card, so:

$Q_B = f(P_B, Q_A)$ (2).

Should we label this market "two-sided" because each side exerts an externality on the other? Is it inevitable that one side will require a subsidy from the other to achieve efficiency?

Economic literature provides an unclear definition for a "two-sided" market. For example, Julian Wright defines a market as two-sided if there are "two distinct types of users, each of whom obtains value from interacting with users of the opposite type over a common platform."[17] According to this definition, however, virtually every market qualifies as a two-sided one—even fisherman and fish buyers benefit from interacting with one another at the fish market, itself a "platform."

Presumably, Wright's definition implies that Equations 1 and 2 above illustrate a two-way externality. However, this definition ignores Coase's insight that, absent transaction costs, negotiations will eliminate any externality.[18] Rochet and Tirole use Coase's principle to restrict "two-sided"

[17] Julian Wright, *One-Sided Logic in Two-Sided Markets*, 3 REV. NETWORK ECON. 44, 44 (2004).

[18] *See* Coase, *supra* note 9, at 1-15.

markets to those in which there are externalities between Markets A and B, as in Equations 1 and 2, and where transaction costs are sufficiently high to prevent the elimination of the externality between individual agents.[19] They further add the condition that a "two-sided" market possess the property that a per unit fee (e.g., a tax) imposed on one side of the market is not neutral—had the same fee been imposed on the other side of the market, the resulting equilibrium would change.[20] In the context of payment systems, an interchange fee functions much like a combined tax and subsidy, with the tax falling on cash and card customers alike, and the "tax revenue" being used to subsidize card issuers and their cardholders. According to Rochet and Tirole, the market must be two-sided for the interchange fee to have an effect.[21]

Thus, if the allocation of costs between the sides of a market does not affect the outcome, then under the Rochet and Tirole definition, the market is one-sided. For example, assume there is a "platform" transaction cost of $1 per item. In a competitive market with uniform transaction costs across consumers and firms, it is immaterial whether the seller or the buyer pays the $1 because the product price will adjust such that the net price paid by buyers and the net amount received by sellers will remain constant. The market is said to be "neutral" with respect to the allocation of the charge because the two groups of end users—buyers and sellers—interact with one another through the prices of the goods or services exchanged.

Economics literature explains that businesses may allocate the costs of selling goods in a two-sided market in a variety of ways.[22] For example, consider Microsoft's sale of its operating system for computers. The usefulness of the

[19] Rochet & Tirole, *supra* note 1, at 9-16.

[20] *Id.*

[21] *Id.*

[22] The very definition of a two-sided market, according to Rochet and Tirole, is that changes in the allocation of charges between the two sides are not neutral. *Id.* at 2.

operating system to consumers depends on the number of application programs that use the operating system, but the demand for application programs depends on the number of consumers using the operating system. Microsoft could collect revenue from the sales of its operating system, sales of its applications programs, or the sales of instructions that allow others to write application programs. It could also subsidize any of these activities. How a seller chooses to price its various products to maximize its profits will depend on the interactions among the products.

When a market is two-sided, revenues can come from both sides of the market. Holding constant the sum of the two prices, the revenues collected will depend upon exactly how much each side of the market is charged.[23] Consider the similar public finance problem of raising sufficient tax revenue to cover an expenditure. Theories on optimal taxation explain how to tax efficiently each agent. In a two-sided market, the optimal price to charge each side of the market depends on demand elasticities (and cross-elasticities) and externality effects. Economics literature stresses the difficulty inherent in determining the amount that each side ultimately pays; one group might receive a net subsidy, whereas the other group might make net payments.[24] Therefore, in a two-sided market, fee allocation among the various groups becomes an interesting problem similar to optimal taxation. This allocation issue is separate from (though related to) that of determining the total fee amount. Most of the literature on two-sided markets accords with the optimal taxation literature on Ramsey pricing and focuses on linear prices (taxes)—however, this is only a

[23] If a market is one-sided, total revenue will depend only on the total fee per transaction, not on its allocation among market participants.

[24] Rochet and Tirole cite newspapers and TV programs as instances where one side (readers or viewers) receives the product for free while the other side (advertisers) makes net payments. *See* Rochet & Tirole, *supra* note 1, at 34.

convenient simplification.[25] More generally, one can think of applying non-linear taxes (e.g., a lump sum tax plus a linear tax) to each market participant, in an effort to raise revenue in the most efficient manner, while also considering transaction costs and externalities affecting each agent.

A one-sided market potentially can be converted into a two-sided market through the imposition of lump sum taxes. This could be accomplished, for example, if revenue is collected from one group in the form of a fixed fee and used to subsidize (e.g., in amounts proportional to usage) the other group. A fixed charge that is less than the total surplus obtained by a member of the first group will have no effect on usage, while a proportional rebate on the other group can stimulate usage. In a two-sided market, shifting revenue generation from one side to the other is, by definition, possible because participants cannot use side payments to negate the effect of this shifting.

Economic theory of two-sided markets proceeds from a set of assumptions about externalities and transaction costs to the conclusion that efficiency requires that prices reflect the demand interactions between market participants from different groups. Economic theory does not indicate, however, whether these externalities and interdependencies are empirically significant, or whether they should alter how a particular industry is typically analyzed. Rochet and Tirole note that, in general, a firm is a two-sided "platform"—yet, they seem to suggest that this observation is not necessarily critical to economic analysis.[26] In the next section, we describe fees imposed by credit card networks and explain how they fit into the two-sided market framework. Later, we discuss a series of questions raised by the use of that framework to defend the efficiency and

[25] *See, e.g.*, Bruno Jullien, *Two-Sided Markets and Electronic Intermediaries* 12-13 (CESifo, Working Paper No. 1345, 2004), *available at* http://www.cesifo.de/~DocCIDL/1345.pdf.

[26] Rochet & Tirole, *supra* note 1, at 13 ("In competitive environments, firms are often *de facto* one-sided platforms, in that there is little 'wriggle room' for them to manipulate the price structure").

appropriateness of the networks' policies, such as the setting of interchange fees.

III. COLLECTIVE RESTRICTIONS AND INTERCHANGE FEES IN TWO-SIDED PAYMENT SYSTEMS

Confusion sometimes persists over how interchange fees and other network fees relate to the costs incurred in payment systems. Two possible costs exist: (1) the network cost, i.e., the cost of operating a central electronic network; and, more interestingly, (2) the costs card-issuing banks incur to promote their cards and to serve their customers. We begin by reviewing network fees designed to cover network costs and then discuss interchange fees and their purported rationale. We conclude with a series of questions designed to probe the procompetitive justifications offered for certain collective policies, such as interchange fees, where these justifications are based on the two-sided nature of the market.

A. Fees to Cover Network Costs

MasterCard and Visa represent large associations of one or both of the following types of financial institutions: "issuers" (those that issue credit and debit cards) and "acquirers" (merchants that accept those cards). The associations incur costs primarily to build and to operate their networks (i.e., the processing and settlement systems that they operate directly), rather than merely to act as standard-setting organizations. Card systems fund most of these costs by charging fees to their issuing and acquiring members.[27] Because both a merchant and a cardholder (and

[27] These fees are paid to and retained by the network and are distinct from the interchange fee, which is a payment administered by the network and made from the merchant's bank to the cardholder's bank. *See, e.g.*, MasterCard Inc., Annual Report (Form 10-K), at 2 (Dec. 31, 2004), *available at* http://www.mastercardinternational.com/corporate/corp_stat.html. ("Generally, interchange fees are paid by acquirers to issuers in connection with transactions initiated with our cards. These fees

their respective banks) participate in each transaction, there are two "sides" of each transaction that can be charged fees to fund the association's costs. If these payment markets are two-sided (using the Rochet and Tirole definition), then the division of the association's own costs between the two sides will have real allocative effects.

For the sake of simplicity, suppose that credit cards are the only means of payment (as is largely the case with respect to some Internet merchants). To cover the network costs, Visa (and its bank members) could charge a fee to the merchant and/or the purchaser. The final equilibrium could depend on how the fees are allocated between the two sides.[28] If so, the market is two-sided, and the allocation of network fees between the two sides affects the outcome.

Now suppose that cash represents another payment method.[29] Assume further that transaction costs preclude the market from charging different prices for cash and credit purchases. In that case, imposing a fee on the merchant will affect the prices the merchant charges each cash and credit customer, depending upon whether the network system was initially charging cardholders or merchants directly. Cash customers would face lower prices if cardholders bore all the fees (e.g., through a charge appearing at the end of the month on their credit card statements) than if merchants were to bear them and pass them on to *all* customers by raising retail prices. Hence, in this scenario, the market is two-sided because of the transaction costs that prevent cash and credit retail prices from differing.

We are unaware of any current legal controversies specifically regarding the relative allocation of MasterCard

are collected from acquirers and passed to issuers . . . We administer the collection of [interchange fees] through the settlement process; however, we generally do not earn revenues from them.").

[28] For example, this could occur if consumers pay less attention to fees that appear on their bank statement than they do to prices posted by the retailer.

[29] For simplicity, we use "cash" to represent cash as well as other means of payment that are less costly to the merchant than credit cards, such as checks and PIN debit cards.

or Visa network fees (as distinct from interchange fees) between issuers and acquirers.[30] One reason may be that the card associations' network costs constitute a relatively small fraction of total sales volume. MasterCard's total costs, for example, apparently amount to only about 13¢ per transaction, an amount equal to less than 0.19% of the average purchase transaction amount of $72.28.[31] In practice, card associations cover a significant fraction of their own network costs due to acquirers by charging merchants.

B. The Interchange Fee

In addition to the fees MasterCard and Visa charge members to cover network costs, each association also requires that, in connection with each transaction, the merchant's acquiring bank (and thus, in practice, the merchant) pay an "interchange fee" to the cardholder's issuing bank. The interchange fee greatly exceeds the magnitude of fees collected by the associations themselves to operate their networks. One analyst recently estimated that the weighted-average interchange fee on MasterCard and Visa card transactions is 1.75%—and rising.[32] The level of credit card interchange fees is thus an order of magnitude greater than the associations' own costs and sufficient to have significant allocation effects.

Because interchange fee revenue flows directly to card issuers and greatly exceeds the transaction fees that the

[30] There have been controversies, however, over the manner in which the associations collect and disclose fees to finance their own network in connection with foreign currency transactions. In addition, First Data—a firm that processes transactions on behalf of bank members of MasterCard and Visa—has alleged in an antitrust action that Visa violated the antitrust laws when it "required First Data's customers to pay for VISA's network services regardless of whether the customers use those services" and because Visa "uses its fee structure to raise the costs of rivals and potential rivals . . . ." First Data Co., Quarterly Report (Form 10-Q), at 59 (September 30, 2002).

[31] Annual Report, *supra* note 27, at 45.

[32] Ken Posner, *The Empire Strikes Back*, MORGAN STANLEY EQUITY RES. (Morgan Stanley, New York), Mar. 8, 2005, at 3.

issuers pay to MasterCard or Visa, the net effective price charged to an issuing bank for a transaction is negative: On average, the card-issuing bank *receives* net fee payments when one of its cardholders makes a purchase with its card. In addition, many issuers offer additional perks, rebates, and incentive benefits based on card usage; these perks are frequently provided even to cardholders who never pay any fees.

Thus, interchange fees are not neutral, but instead have real allocation effects. This is true partially because in many cases there is a pooling equilibrium in which cash customers pay the same retail price as credit card customers. If the interchange fee rises, then merchants raise retail prices to *all* customers. The interchange fee, therefore, has a real effect because higher retail prices reduce purchases by cash customers. If merchants instead charged retail consumers prices reflecting the merchant's differential costs of cash versus credit, then an increase in the interchange fee would have no effect on cash customers.[33] Increasing interchange fees would affect credit card customers, but the issuing bank could offset the higher price with a rebate resulting from the funds the issuing bank receives from the increased interchange fee. If so, the market would be neutral with respect to the interchange fee.

Merchants, however, tend not to charge different prices based on modest differences in payment method costs because of the transaction costs that would likely arise from administering multi-tiered pricing (and at least sometimes because of the existence of legal restrictions).[34] Moreover, MasterCard and Visa enforce "no surcharge" and "no discrimination" rules against merchants to reinforce a pooling equilibrium in which customers pay the same price

[33] Alternatively, cash and credit card customers could pay the same retail price, but the credit card customer would be charged the interchange fee that appears on monthly credit card statements.

[34] *See* Alan S. Frankel, *Monopoly and Competition in the Supply and Exchange of Money*, 66 ANTITRUST L.J. 313, 341-49 (1998).

irrespective of payment method.[35] The important consequence of this pooling equilibrium is that the interchange fee imposes a tax on cash customers. For example, if the interchange fee rises by $1, and half of a merchant's customers pay with cash and the other half pay with credit cards, then the merchant would have to raise retail prices to all customers by $0.50 to cover its increased costs.

The economic controversies over interchange fees center on whether the real allocation effects generated by interchange fees and associated vertical restrictions on pricing *correct* an underlying externality that otherwise would result in inefficiency or rather *cause* such inefficiencies (and wealth transfers). The remainder of this Section describes how these controversies relate to the two-sided market framework and poses some of the questions raised by MasterCard's and Visa's collectively set policies.

C. Usage and Membership Externalities

Baxter first described interchange fees as the solution to the externality induced by transaction costs.[36] Baxter

[35] For example, MasterCard rules provide that "[a] merchant must not directly or indirectly require any MasterCard cardholder to pay a surcharge or any part of any merchant discount or any contemporaneous finance charge in connection with a MasterCard card transaction." MASTERCARD INT'L INC., MERCHANT RULES MANUAL § 9.12 (2005), *available at* http://www.mastercardmerchant.com/docs/accept_mastercard/merchant_rules.pdf. Moreover, "[a] merchant must not engage in any acceptance practice that discriminates against or discourages the use of MasterCard cards in favor of any other acceptance brand." *Id.* Visa has similar rules. A merchant can charge lower prices for cash customers, but it cannot charge differentially among customers who use different brands of cards.

[36] Baxter, *supra* note 1, at 553 n.9 ("In four-party payment mechanisms . . . a side payment between [consumer] and [merchant], coupled with payment by each [of them] to [issuer and acquirer], respectively, in amounts equal to respective bank costs but not to respective marginal utilities of [consumer] and [merchant], is theoretically sufficient to attain equilibrium. That in practice side payments between banks occur instead is strong evidence that higher transaction costs

postulated that credit card payments saved significant transaction costs over cash and check payments, and merchants should therefore charge lower prices to credit card customers in order to encourage use of the cards.[37] Recall the earlier example of how airlines incentivize consumers to use the Internet to purchase tickets. But, unlike the current airline ticket market, a pooling equilibrium persists in retail payment systems because merchants find it costly to charge consumers different retail prices depending on payment method they use.[38]

Baxter concluded that inadequate incentives would exist for consumers to pay with credit cards because transaction costs make it unprofitable for merchants to offer lower retail prices for credit purchases. Baxter postulated that an efficient interchange fee, one that could effectively replicate the discount merchants would levy on credit purchases if merchants could costlessly implement such pricing, would solve this problem.[39] Card associations effectively transmit this amount from the merchant to the issuing bank, which then rebates the fee to the cardholder—assuming the bank operates in a perfectly competitive market. For example, if the merchant wanted to charge cash customers $10 and credit card customers $9 but could not set different retail prices due to transaction costs, then he could replicate that outcome by setting the retail price to all customers at $10, paying an interchange fee of $1 to the card-issuing bank, and having the issuing bank rebate the interchange fee to the credit card customer.

The Baxter inefficiency and proposed efficient interchange fee have little to do with common descriptions of network effects in payment systems. The interchange fee, as

characterize side payments that take the form of price adjustments between the principals.").

[37] *Id.*

[38] *See* Frankel, *supra* note 34, at 314. In fact, as already described, the associations have rules that limit or prevent merchants from charging surcharges to consumers who use the associations' cards to complete transactions.

[39] *See* Baxter, *supra* note 1, at 553-56.

Baxter describes, does not correct a two-sided membership externality, but rather a *usage* externality.[40] This problem is distinct from the "chicken and egg" problem described earlier, and it exists irrespective of network size. In this setting, the interchange fee is used to replicate the pricing that would prevail in a world of zero transaction costs.

In addition to Baxter's efficiency justification for an interchange fee, two-sided market literature propounds several justifications for collective action related to interchange fees. Those justifications typically rely on economic models in which externalities are assumed to exist and are those that can be corrected by an interchange fee. For example, Rochet and Tirole assume that individual card issuers possess market power.[41] The consequence of market power is that there is too little output of credit cards. Rochet and Tirole show that an interchange fee can correct this problem by paying issuing banks to expand output.[42] The source of funds, however, is a tax on cash customers. In addition, the interchange fees are said to "balance" merchant acceptance with adequate incentives for issuers to promote their cards at prices low enough to induce cardholder usage.

D. Policy Questions Regarding Interchange Fees and Associated Pricing Restrictions

As a general matter, it is prudent to carefully approach intervention in the operation of any firm. MasterCard and Visa, on the other hand, are joint ventures whose members

[40] Baxter himself does not use the term "usage externality," but subsequent authors (such as Rochet and Tirole) have used it to describe Baxter's insight that the customer's choice of payment mechanism affects merchant costs without those costs being reflected in the price to the consumer. Rochet and Tirole recognize the distinction between usage and membership externalities and analyze both of these effects. *See* Rochet & Tirole, *supra* note 1, at 20.

[41] Jean-Charles Rochet & Jean Tirole, *Cooperation Among Competitors: Some Economics of Payment Card Associations*, 33 RAND J. ECON. 549, 552 (2002); *see also* Richard Schmalensee, *Payment Systems and Interchange Fees*, 50 J. INDUS. ECON. 103 (2002).

[42] Rochet & Tirole, *supra* note 41, at 559.

have collective market power. Therefore, one must ask whether collective actions concerning interchange fees and other rules that affect retail prices are in the public interest, or whether they are the actions of a cartel that harm the public. In this section, we pose a series of questions that probe the validity of some of the efficiency defenses offered for imposing interchange fees in two-sided markets.

1. *If a theoretical model shows that it is possible that a collectively fixed price (in this case, interchange fee) might improve economic efficiency, should that be enough to justify the practice in the context of a two-sided market, even when that rationale would likely fail in a standard price-fixing case?* For example, Rochet and Tirole show that interchange fees might achieve efficiency by offsetting the output-reducing effects of unilateral market power exercised by card issuers. Their model accords credit card issuers with market power, and hence, suggests that too few cards are used.[43] To offset that output restriction, interchange fees subsidize the issuance of credit cards, thereby acting like a tax on cash purchases. Suppose instead that credit card issuers exercised a cartel within the salt industry (i.e., collected a tax on salt) and used the resulting profits to subsidize credit card issuance, or that the government imposed a tax on salt and used the tax revenues to subsidize credit card issuance. If it is appropriate to allow credit card issuers to use the interchange fee to tax cash customers in order to expand output, would it also be appropriate to allow them to cartelize the salt industry to achieve the same purpose?

2. *Does the existence of pricing restrictions on merchants refute Baxter's justifications for interchange fees?* Suppose Baxter is correct in contending that interchange fees are required to enable merchants to charge a lower effective price to credit card customers. In such a case, merchants would not require a no-surcharge rule—a rule preventing merchants from charging a higher price for credit cards—because absent transaction costs, merchants would want to charge a lower, not higher, price to credit card customers.

[43] *Id.* at 552.

Moreover, if an interchange fee is necessary because merchants are supposedly unable to charge cash and credit customers different retail prices, then it seems odd to prohibit merchants from charging whatever prices they want. Indeed, if merchants differ from each other, the "optimal" interchange fee will differ from merchant to merchant, and one would think that Baxter's argument would permit those merchants who can set different retail prices to do so.

3. Does the "chicken and egg" problem represent a network externality, and if so, should it matter to antitrust policy? The "chicken and egg" network externality problem arises when externalities are so extensive that there is an inadequate incentive for a firm to enter the market. Credit cards, however, have achieved widespread consumer adoption and almost universal acceptance among major retailers. It therefore remains difficult to defend the present need for collective setting of interchange fees on the grounds that the fees were necessary to launch the network decades earlier.

4. Is there a significant free-rider problem among issuers? Some economists defend the utility of vertical contractual practices, like exclusive territories and resale price maintenance, to combat free-riding problems that otherwise make it unprofitable for individual distributors to offer promotional services that consumers value.[44] Is there an analogy between this procompetitive justification for vertical restrictions, interchange fees, and associated pricing restrictions?

Arguably, when one issuer increases the number of cardholders, it induces more merchants to accept cards, which in turn increases the demand for cards issued by all member banks. If the market is already highly penetrated, however, then, at the margin, there may be little external benefit from expanding the number of cards held or by increasing the number of merchants accepting cards.

[44] DENNIS W. CARLTON & JEFFREY M. PERLOFF, MODERN INDUSTRIAL ORGANIZATION 437 (4th ed. 2004).

Moreover, if the point of interchange fees is to induce issuers somehow to seek out additional card users, why not simply allocate the "interchange fee" directly to the card customers (e.g., as a fee appearing on monthly credit card account statements), and omit all charges to the merchant? Such a mechanism would avoid the "tax" on cash customers, while still preserving the flow of revenue to the issuing bank. The failure of the market to structure the fee in this manner suggests that taxing cash consumers encourages the existence of interchange fees and their ability to "balance" demand in the two-sided market.

5. *Is antitrust regulation a more appropriate forum for analyzing interchange fees and vertical restrictions?* Some countries have investigated interchange fees, no-surcharge rules, and other restrictions in payment markets as issues arising under antitrust or competition law, while central banks have historically taken a more regulatory approach.[45] Is one approach better suited than the other to addressing these issues? In addition to antitrust issues, do interchange fees raise other significant public policy concerns, and if so, does resolution of those other concerns conflict with competition analysis or reinforce those conclusions? For example, low income and minority households are far more likely to use cash exclusively than are more affluent households: only 28.5% of families with annual income below $10,000 possess a bank credit card, compared to 95.8% of families with incomes above $100,000,[46] and only "59% of African-American households had credit cards in 2001, compared to 53% for Latinos and 82% for whites."[47] The

[45] Stuart Weiner & Julian Wright, *Interchange Fees in Various Countries: Developments and Determinants* 1 (Fed. Res. Bank of Kansas City, Working Paper No. 05-01, 2005), *available at* http://www.kc.frb.org/FRFS/PSR/WeinerWrightWorkingPaper101205.pdf.

[46] *See* U.S. CENSUS BUREAU, STATISTICAL ABSTRACT OF THE UNITED STATES: 2004-2005 t.1191, *available at* http://www.census.gov/prod/2004pubs/03statab/banking.pdf.

[47] *See Study Shows Card Use Linked To Race*, CARDLINE, May 24, 2005, *available at* http://www.cardline.com (citing a study based on data

interchange fee, therefore, may disproportionately harm minorities and the poor because it acts as a tax on cash customers.

6. *Should the existence of proprietary systems like Discover Card and American Express affect the applicability of antitrust regulation to payment joint ventures such as MasterCard and Visa?* If a single firm, as opposed to a joint venture, employs its credit cards as part of a proprietary system—issuing cards to consumers and acquiring transactions from merchants—it does not need to set an interchange fee. For example, American Express, unlike Visa, does not charge interchange fees, but rather sets a merchant discount;[48] this way, it may charge cardholders and promote its cards as it sees fit. Preventing MasterCard or Visa from charging interchange fees could place them at a competitive disadvantage to American Express because they would face constraints that their integrated rival does not. Does that justify allowing bank members of MasterCard and Visa the unfettered ability to collectively set interchange fees? Or, does it suggest the opposite: that constraints on MasterCard and Visa interchange fees actually restrict the amount that American Express can charge?

IV. CONCLUSION

In our discussion, we have suggested answers to some of these questions that may help form the basis for informed decisionmaking.[49] Nonetheless, it is uncertain whether any

from the Federal Reserve's Survey of Consumer Finances for the years 1992 through 2001).

[48] Rochet & Tirole, *supra* note 41, at 563. This statement assumes that American Express remains exclusively a proprietary system that both issues and acquires.

[49] *See also* Dennis W. Carlton & Alan S. Frankel, *The Antitrust Economics of Credit Card Networks*, 63 ANTITRUST L.J. 643 (1995); Dennis W. Carlton & Alan S. Frankel, *The Antitrust Economics of Credit Card Networks: Reply to Evans and Schmalensee Comment*, 63 ANTITRUST L.J. 903 (1995); Dennis W. Carlton & Alan S. Frankel, *Antitrust and Payment Technologies*, 77 FED. RES. BANK OF ST. LOUIS REV. 41 (1995); Frankel, *supra* note 34; Alan S. Frankel, *Interchange Fees in Various Countries:*

consensus will emerge on these contentious issues. In light of the different approaches taken in other countries, a body of evidence will likely develop to confront the highly theoretical economic models offered in defense of interchange fees, the no-surcharge rule, and other controversial practices in two-sided retail payment systems.[50]

Comment on Weiner and Wright (Fed. Res. Bank of Kansas City, Presented at Interchange Fees in Credit and Debit Card Industries: What Role for Public Authorities?), (May 5, 2005), *available at* http://www.kansascityfed.org/FRFS/PSR/PDF/FrankelDiscussantRemarks_final6-8-05.pdf; Alan S. Frankel & Allan L. Shampine, *House of Cards: A Competitive Analysis of Interchange Fees*, ANTITRUST L.J. (forthcoming).

[50] Australia's central bank, for example, has significantly reduced the level of credit card interchange fees and has suggested that further reductions, perhaps towards a par collection system, might be forthcoming. *See* RESERVE BANK OF AUSTRALIA, REFORM OF THE EFTPOS AND VISA DEBIT SYSTEMS IN AUSTRALIA, A CONSULTATION DOCUMENT 2 (2005), *available at* http://www.rba.gov.au/PaymentsSystem/Reforms/Eftpos/ConsultDocFeb2005/.