Winter 2003

The Payment Card Economics Review

Volume 1

Two-Sided Markets and Interchange Fees

An occasional publication of payingwithplastic.org.

TABLE OF CONTENTS

INTRODUCTION

This volume is about interchange fees. A once obscure aspect of checks, credit cards and other systems for exchanging value among individuals and businesses, interchange fees are now of considerable interest to economists, lawyers and competition regulators around the world.

All payment systems involve two sets of customers—those who offer payment and those who accept payment—and an entity that facilitates the exchange of value between those two sides. Some are three-party systems. The entity stands between the payor and payee. American Express reimburses the merchants who accept its cards and collects money from the individuals who use its cards. Other systems are four-party (or perhaps more accurately five-party): an issuer collects money from its cardholders, an acquirer collects money from merchants it services, and a system coordinates the transfer of payment between the issuers and acquirers that belong to its system.

Coordination is easy in three-party systems. One hand of American Express takes the money from merchants (having added a charge for its services) and gives it to the other hand of American Express for collecting from cardholders (perhaps adding in late fees or finance charges depending on the card type). Coordination is more complex in four-party systems. Does the acquirer turn over all the money the payee received to the issuer? Or does it keep some for itself—after all, the issuer needed the acquirer for the transaction to happen? Or does it turn over all the money and pay a fee to the issuer—the acquirer needed the issuer for the transaction to happen? Should the acquirers and issuers just negotiate with each other, or should the system set a rule? And what should that rule be?

Interchange is the fee that the issuer or acquirer pays each other. Almost all payment systems have rules for setting this fee. In the United States banking regulations long ago resulted in checks being exchanged at par—that means the interchange fee is zero. In most countries, associations of financial institutions that operate debit and credit card systems have rules that require the acquirer to pay the issuer a percent of the transaction. The rules for ATM sys-

tems vary. On ATM transactions, the issuer tends to pay the acquirer, but the reverse is generally true for debit transactions.

Interchange fees have generated controversy—they can be interpreted as a price set by competitors. Most payment card systems are run by associations. Members compete with each other for cardholders and merchants. The association does not set the prices the members charge to these customers. This is apparent in the United States, especially, where there is highly visible competition for cardholders—registered in mailings, advertisements for free cards, and low interest rates on credit cards—and less visible but still intense competition for merchants. The members cooperate in operating a network for processing transactions—this includes the "railroad" for moving money, a brand name, and rules for managing the complex interplay among members and the system. This is true for the global payment associations—Visa and MasterCard—as well as national associations such as Cartes Bancaires in France. The members, working through the association, cooperate on a rule for how much acquirers have to pay issuers for each transaction—that is, they "fix" the interchange fee.

The associations have defended that rule as based on necessity and efficiency. Necessity: there are too many combinations of issuers and acquirers for them to contract with each other, and too many possibilities for opportunism and free-riding. Efficiency: the interchange fee is necessary for balancing the relative demands of cardholders and merchants and optimizing the value of the system. Opponents—private parties, merchant associations, or competition agencies—argue that the interchange fee is just a collective price fix and just as bad as any other cartel price. Moreover, they say, the interchange fee gets passed on to the merchant; for various reasons the merchant passes the cost on to all customers rather than just to cardholders. The result is that cash and check-paying customers end up subsidizing card-paying customers.

A U.S. appeals court sided with the associations in a decision in the 1980s—the Supreme Court declined to consider the decision—and as a result the interchange fee is recognized as a procompetitive device in the United States. The Reserve Bank of Australia (RBA) concluded that a fix is a fix—since the associations could not establish that the interchange fee they set was the best

one from the public's standpoint, the RBA should set the fee instead of the associations. That decision is currently being appealed. The European Commission also went the regulatory route—in a settlement, they agreed to permit the fee only if Visa lowered it and kept it under certain cost-based benchmarks.

Although interchange fees are unique to payment systems, the set of economic problems addressed in this volume are not. Payment card systems are an example of a two-sided market, one in which there are two classes of customers who need each other for a product to even exist. Video game console developers need game developers and game users. Operating systems are similar: vendors from BeOS to Microsoft to Palm need people who use the operating system and software developers to write applications. Exchanges from the FTSE to e-Bay need buyers and sellers. Advertising-supported media need to figure out ways to match advertisers with readers. Payment card systems need to have customers who pay with cards and merchants who take those cards. In two-sided markets businesses have to get both sides on board. The economic analysis of interchange fees has resulted in a deeper understanding of pricing and investment strategies in these kinds of industries, as well as the competitive consequences of those strategies. The interchange fee is part of a general class of strategies used in two-sided markets to balance the demands of the two sides of the market.

This collection of essays is designed to broaden perspectives about interchange fees as well as provide an extensive case study of a feature of a globally important two-sided market. David Evans presents a non-technical introduction to the economics of two-sided markets in the first essay. After summarizing the origin and role of interchange fees, Howard Chang contrasts the approaches taken by the United States and Australia towards evaluating whether the interchange fee is in the public interest. Richard Schmalensee, in the third essay, then summarizes the economic literature on interchange fees—from William Baxter's pathbreaking work in the early 1980s to the recent theoretical advances by Jean-Charles Rochet and Jean Tirole. The remaining essays consist of articles that have been published (or are forthcoming) in peer-reviewed journals in economics or the law.

IT TAKES TWO TO TANGO: THE ECONOMICS OF TWO-SIDED MARKETS

David S. Evans,
NERA Economic Consulting

Japan's dating clubs—typically bars or cafes—offer startling ways to meet the opposite sex.[2] At one club, men and women sit on opposite sides of a glass divide. If a man sees a woman he likes, he can ask a waiter to carry a "love note" to her.

But it takes two to tango: enough men must participate to attract women, and enough women must show up to engage the men. The club must thus figure out how much to charge each of the sexes to get the right mix of patrons while still generating profits for the owners. One bar charges men $100 for membership plus $20 a visit, and lets women in for free. A pricing structure that obtains a disproportionate share of the revenues from men is common in singles bars, discotheques and other enterprises that are effectively in the matchmaking business.

A. WHAT ARE TWO-SIDED MARKETS?

Dating clubs are one example of a "two-sided" market, in which there are two classes of customers, and each type of customer values the service only if the other also buys the service. Indeed, in such markets the product or service only has value when it is consumed jointly.

Two-sided markets generate positive "externalities" by bringing the other side on board (lots of guys to meet). For that matter, two-sided markets only exist because of the inability of the two sides to internalize these externalities without an intermediary. Firms generate benefits for themselves (in the form of profits) and for society in general by figuring ways to internalize these externalities.

Many high-profile industries, including some that are central to the technologically based new economy, are grounded on business models similar to those of dating clubs. Consider these examples:

- Computer operating systems provide features that software developer can use in creating applications, along with the platform on which computer users can run the applications. Both software developers and users are needed for the operating system to be a viable product: the success of the Palm OS for handheld devices, Microsoft Windows for the desktop, and Sun Solaris for server computers all depend on attracting large numbers of customers on each side of the market.

---

1 For a more detailed discussion of the material presented in this paper see David S. Evans, *The Antitrust Economics of Two-Sided Markets*, AEI Brookings Related Publication 02-13 (Sept. 2002) (visited Oct. 21, 2002) <http://aei.brookings.org/admin/pdffiles/phpMt.pdf>.

2 Howard W. French, *Osaka Journal; Japanese Date Clubs Take the Muss Out of Mating*, N.Y. TIMES, Feb. 13, 2001.

- Video games have a parallel symbiotic relationship with proprietary game consoles such as the Sony PlayStation. Game developers have strong financial incentives to write for consoles that attract lots of players, while game enthusiasts will only buy consoles with lots of games to choose from. Thus, console manufacturers must lure both developers and users.

- Payment card systems—credit, debit and charge cards—are yet another example. Consumers use them to make payments, while retailers use them to take payments. Merchants are more willing to accept cards that are more widely held by shoppers, and shoppers are more willing to carry cards that are widely accepted by merchants.

- Industries that "make markets" by arranging for buyers and sellers to meet each other are also two-sided markets: Internet-based business-to-business exchanges, real estate brokers, and corporate bond exchanges are but a few examples.

To succeed, any business in a two-sided market must create a pricing structure that brings balanced numbers to each side of the table. And strategies differ along with the factors that affect each side of the market differently.

Most computer operating system vendors do not seek significant revenues from software developers, choosing instead to collect from users of the operating system or from the sale of complementary hardware (such as the Palm organizers and Sun server computers). Sellers of video game consoles, by contrast, do earn significant revenues from the game developers. Charge cards, such as American Express's, earn a disproportionate share of their revenue from merchants. For their part, print media such as magazines and newspapers typically give readers content for a fraction of the cost of the service, collecting the bulk of their revenue from advertisers.

Note the key distinction here: unlike firms in traditional industries, those in two-sided markets must worry about the price *structure* as well as the price *level*. In two-sided markets, the product may not exist at all if the business does not get the price structure right.

Most, if not all, industries characterized by "network effects"—where the value of a product to each user increases with the total number of users—are two-sided markets. Think of the fax machine: you only value the machine if there are a lot of people to whom you can send faxes and who can send faxes to you. Indeed, network effects usually arise because the product is two-sided – a point that is obvious when there are two distinct types of customers, such as men and women in the dating club example.

Both two-sided markets and markets characterized by network effects raise novel questions about the workings of competition, and thus have attracted the interest of American and foreign antitrust enforcement agencies. Indeed, businesses that compete in two-sided markets have figured prominently in a variety of high-profile cases in the last decade:

- the AOL-Time Warner merger, where U.S. and European authorities investigated the impact on two-sided markets including Internet portals, magazines and free television;

- the credit card association investigations, where Australian and European authorities examined two-sided markets involving merchants and card users;

- the American, European and private antitrust cases against Intel, which competes in a two-sided computer hardware platform market;

- the Microsoft competition cases, where U.S. and European authorities investigated two-sided markets involving operating systems and software that might emerge as alternative computer platforms; and

- the probes into online securities broker-dealers, where six separate U.S. regulatory investigations and one European investigation are investigating possible anticompetitive behavior.

In some cases, the two-sided nature of the market is central to allegations of anticompetitive behavior. For example, the credit card investigations focused on the pricing structure used to balance two-sided demand, while *U.S. v. Microsoft* included the claim that Microsoft's strength on one side of the market (applications software) was the source of a barrier to entry to the operating system business. In other cases, the two-sided nature of the market provided an important backdrop for understanding the workings of the industry.

B. THE ECONOMICS OF TWO-SIDED MARKETS

A market is two-sided if at any moment (a) there are two distinct groups of customers, (b) the value obtained by one kind of customers increases with the number of the other kind of customers, and (c) an intermediary is needed to internalize the externalities created by one group for the other group. Two-sided markets are typically served by businesses that supply both sides and that adopt pricing and investment strategies tailored to getting—and keeping—both sides on board.

Jean-Charles Rochet and Jean Tirole have shown that firms in two-sided markets have to choose a *pricing structure* as well as a *pricing level* to maximize profits.[3] The pricing structure determines the *relative* prices charged on the two sides of the market—that is what men pay relative to women, software developers relative to software users, cardholders versus merchants. The optimal-structure depends on the elasticities of demand and the marginal cost of pro-

3 Jean-Charles Rochet & Jean Tirole, *Platform Competition in Two-Sided Markets*, J. EUR. ECON. ASS'N (forthcoming).

viding services on both sides of the market. When properly set, the pricing structure marshals enough demand from both sides to make each side value the product.

None of the formal conditions for determining the price level or the price structure in models of two-sided markets corresponds to equating marginal revenue with marginal cost on either side of the market. In fact, such conditions have no meaning in two-sided markets, because there is no conceptual way to allocate the increases in revenues from changes in prices to one side or the other. Changes in prices result in more "transactions" from which both sides benefit. Nor is there any useful way to allocate the costs. Often costs are jointly incurred, and any means of allocating them is arbitrary. These results are broadly similar whether the seller is a monopolist, or one of many competing firms selling to both sides of the market.

In practice, consumers in two-sided markets tend to engage in *multihoming*—that is, consumers on at least one side of the market rely on more than one seller of services. For example, game developers write for several consoles, merchants accept several brands of credit cards, and homebuyers often use the services of several real-estate agents. Here competing two-sided firms still must choose a price level and a pricing structure. However, the elasticities of demand on both sides of the market are increased by a factor that reflects the extent to which consumers multihome, and therefore have substitutes readily available.

C. BUSINESS MODELS IN TWO-SIDED MARKETS

Although the economics presented above is by necessity simplified, it illuminates the rationale for the business models that have been adopted in two-sided markets. Consider several issues that occur repeatedly in two-sided markets.

1. Getting Both Sides on Board

In two-sided markets, demand on one side vanishes if there is no demand on the other, regardless of how prices are set. Heterosexual men will not go to dating clubs if women do not attend. Merchants will not accept a payment card if none of their customers carry the card. Computer users will not use an operating system if applications software is not available. Sellers of corporate bonds will not use a trading mechanism if buyers won't bid.

One way to get both sides on board is to create a critical mass of users on one side of the market by giving them the service for free, or even by paying them to take it. Diners Club initially gave its charge card away—there was no annual fee, and users got the benefit of the float. Netscape gave away its browser to many users; after Microsoft raised the ante by giving away its browser to all

users, Netscape followed suit. By the same token, Microsoft is reportedly selling its Xbox hardware below cost in order to build a base for game sales.[4]

Another way to solve the chicken-and-egg problem is to invest in one side of the market to lower costs. Microsoft gives away costly tools that help developers to write applications software for Microsoft platforms. Bond dealers take positions in their personal accounts if a bond is thinly traded and the long time delays between buys and sells would hinder the market's pricing and/or liquidity.

Subsidies or transfers to one side of the market helps the platform solve the chicken-and-egg problem by encouraging one group's participation—which in turn encourages the other group's participation. Bernard Caillaud and Bruno Jullien refer to this strategy as "divide-and-conquer."[5] From the perspective of the individual firm, such transfers can yield the added benefit of discouraging patronage of competitors. For example, when Palm provides free tools and support to PDA applications software developers to encourage them to write programs for the Palm operating system platform, it also gives them incentives to invest less in applications for other operating systems.

2. Pricing to Balance Interests

Firms in mature two-sided markets—i.e., those that have already gone through the entry phase in which the focus is on solving the chicken-and-egg problem—still must devise and maintain an optimal pricing structure. Generally, companies settle on pricing structures that are heavily skewed towards one side of the market. For example, in 2001, American Express earned 82 percent of its revenues (excluding finance charge income) from merchants.[6] Microsoft earns the bulk of its revenue from Windows by licensing the operating system to computer manufacturers and retail customers. Real estate brokers in the United States typically earn most or all of their revenues from the sellers.

Sometimes all the platforms converge on the same pricing strategy. Microsoft, Apple, IBM, Palm and other operating system companies could have charged more to applications developers and less to computer users. But they all independently decided that it made sense to charge little or nothing for developers' tools.

---

4 David Becker, *Xbox Drags on Microsoft Profit*, CNET.COM, Jan. 18, 2002 (visited Aug. 21, 2002) <http://news.com.com/2100-1040-818798.html>.

5 *See* Bernard Caillaud & Bruno Jullien, *Chicken & Egg: Competing Matchmakers*, CEPR WORKING PAPER #2885 (Apr. 24, 2001), at 16. *See also* Bruno Jullien, *Competing in Network Industries: Divide and Conquer*, IDEI WORKING PAPER (Jul. 2001), at 1.

6 If finance charge revenues are included, American Express earned 62 percent of its revenues from merchants in 2001. *See* American Express Company Annual Report 2001 (visited Aug. 15, 2002) <http://www.onlineproxy.com/amex/2002/ar/pdf/axp_ar_2001.pdf>, at 35.

With debit cards, by contrast, pricing choices have varied widely. In the late 1980s, the ATM networks had a base of customers who used their cards to withdraw cash or to obtain other services at ATMs; no merchants honored these cards. To add merchant debit services, the ATM networks decided to charge a very modest fee (8 cents on a typical $30 transaction) to merchants. The goal was to convince retailers to install pin-pads that could read the ATM cards consumers already had, and to accept the personal identification numbers they already used to gain access to ATM machines.[7] It worked: the number of pin-pads increased from 53,000 in 1990 to about 3.6 million in 2001.[8]

For their part, the credit card systems had a base of merchants who took their plastic, but they did not offer cards that could be used to debit consumers' checking accounts directly. They imposed much higher fees on merchants than the ATM networks——about 38 cents on a typical $30 transaction.[9] Here, the strategy was to persuade banks to issue debit cards and for cardholders to take these cards, thereby putting pressure on merchants to accept them. The number of Visa debit cards in circulation did, in fact, increase from 7.6 million in 1990 to about 117 million in 2001.[10]

Two other factors influence the pricing structure in two-sided markets. There may be a sub-group of customers on one side of the market—Rochet and Tirole refer to them as "marquee buyers"—who are extremely attractive to customers on the other side of the market. The existence of marquee buyers tends to reduce the price to all buyers and increases it to sellers. A similar phenomenon occurs when some customers are extremely loyal to the two-sided firm—perhaps because of long-term contracts or sunk-cost investments. For example, American Express has been able to charge a relatively high merchant discount as compared to other card brands—especially for their corporate card—because merchants viewed the American Express business clientele as extremely valuable.

Corporate expense clients were thus marquee customers who made it possible for American Express to raise its prices to the merchant side of the market. In the online debit card market, however, card issuers faced "captive" customers:

---

7 *See* DAVID S. EVANS & RICHARD SCHMALENSEE, PAYING WITH PLASTIC: THE DIGITAL REVOLUTION IN BUYING AND BORROWING 300 (1999).

8 *Id.* at 308-309; and THE NILSON REPORT NO. 759 (Mar. 2002), at 6.

9 The ATM systems typically charged a flat interchange fee per transaction, while the interchange fee set by Visa and MasterCard varied with the size of the transaction. The reported interchange fee comparison is from 1998, around the time of substantial growth in debit for the ATM and credit card systems. EVANS & SCHMALENSEE, *supra* note 7, at 300.

10 *See* THE NILSON REPORT NO. 760 (Mar. 2002), at 7; THE NILSON REPORT NO. 500 (May 1991), at 6.

ATM cards could be used as online debit cards, so consumers did not need to be courted to accept the new payment form. Therefore, it has been the merchants—who must install expensive machinery in order to process online debit transactions—who have been courted.

3. Multihoming

Most two-sided markets accommodate several competing two-sided firms, and at least one side usually multihomes. Consider, for example, personal computers, where the two sides consist of PC users and developers of applications. End-users rarely multihome: they employ a single operating system. But developers do multihome. According to Josh Lerner of the Harvard Business School, 68 percent of software firms in the year 2000 developed software for Windows operating systems, 19 percent for Apple operating systems, 48 percent for Unix operating systems (including Linux), and 36 percent and 34 percent for proprietary non-Unix operating systems running on minicomputers and proprietary operating systems running on mainframes, respectively.[11] In fact, in recent years the percentage of software firms developing for non-Microsoft operating systems has increased. The fastest-growing category has been firms creating software Unix operating systems, notably Linux.

Multihoming affects both the price level and structure. Not surprisingly, prices tend to be lower in the presence of multihoming—the availability of substitutes puts pressure on two-sided firms to keep prices down. The seller has more options dealing with a multihomed buyer on the other side, and can steer toward its preferred platform. As buyer multihoming becomes more prevalent, prices to sellers tend to decrease since they have more substitution options.

Even when multihoming is not common, the potential for multihoming may have significant consequences for pricing. The possibility of multihoming may encourage firms to lower their prices on the side of the market in which multihoming could occur. This discourages customers on that side from affiliating with other two-sided firms.

Note, however, that this does not generate a free lunch for all consumers. A seller facing multihoming on one side can charge more to customers on the other side, for whom fewer substitutes are available.

---

11 *See* Josh Lerner, *Did Microsoft Deter Software Innovation?* WORKING PAPER (Jan. 2002) (downloaded Aug. 15, 2002) <http://gsbwww.uchicago.edu/research/workshops/elo/lerner2.pdf>; and CORPORATE TECHNOLOGY DIRECTORY, EDITIONS 1990-2000. The percentages add up to 205, indicating substantial multihoming on the part of developers.

D. TWO-SIDED MARKETS AND SOCIAL WELFARE

Two-sided markets rarely accommodate large numbers of competitors, both because these markets exhibit network economies, and because it is usually expensive to solve the initial chicken-and-egg problem. At least up to a point, larger firms have advantages over smaller firms, because their scale delivers more value in the form of a bigger network. In the case of two-sided markets, larger firms are able to deliver a larger network of customers on one side of the market to customers on the other side of the market. Note, however, that a heterogeneous market base makes it easier to build multiple two-sided firms, because the presence of a wide variety of customers tends to limit the importance of network effects.

Firms in concentrated two-sided markets, like firms in all concentrated markets, may have opportunities to earn supra-competitive profits—that is, profits exceeding the level needed to attract capital to the industry after accounting for risk. Several factors affect the extent to which this can happen.

1. *The degree of competition*. If the competition is sufficiently intense, the losses incurred during the "getting both sides on board" stage of the industry may offset the profits earned during the mature phase. For example, firms entering the payment card industry have all incurred sizeable losses during their startup phases.

2. *First-mover advantages*. In some markets, being first is critical. In others, it may even prove a disadvantage.

3. *The degree of contestability*. Because many of the two-sided markets are fast moving, current leaders often face considerable competition in the form of potential entrants—other platforms striving to displace today's leader.

4. T*he presence of non-profits*. Two-sided markets in which non-profit associations determine the pricing structure are not likely to permit the participants to earn supra-competitive profits. Payment card associations have put what amounts to a non-profit in charge of managing a physical network for members and for determining pricing policies. Pricing is determined by competition among members of the association.

The reality that most two-sided markets support relatively few sellers and exhibit strong network effects raise familiar issues regarding the viability of competition and the logic of government intervention. By the same token, the pricing and investment strategies that firms in two-sided markets use to get both sides on board and to balance demand raise novel ones. Interdependence of demand casts a long shadow over these markets.

Rochet and Tirole make a number of simplifying assumptions that allow comparisons between prices chosen to maximize private interests under a variety of market conditions and the prices that would maximize social welfare. Strikingly, they conclude that a monopoly, a firm with competition, and a

benevolent social planner would adopt similar price structures. Relative prices would differ somewhat. However, they find that prices preferred by firms (monopoly or two-firm oligopoly) would not be biased towards one side of the market compared to the pricing structure that would be adopted by a benevolent social planner. Hence, there is no reason to believe that the direction or magnitude of the cross-subsidies in real-world markets is systematically different from what a wise social planner would choose.

E. CONCLUSIONS

Two-sided markets are becoming increasingly important to the global economy. Firms that provide platforms for multiple customer groups—notably Microsoft in operating systems and Intel in microprocessors—are a critical part of the computer industry. Individual firms and business cooperatives create platforms for merchants and customers to facilitate a large and growing fraction of financial transactions in high-income countries. The increased importance of the Internet for household-to-household, business-to-household and business-to-business transactions, along with the emergence of e-pay systems on the Internet will certainly accelerate this trend. And while it is now plain that the reach of most dot-coms exceeded their grasp, Internet-based businesses are sure to flourish in the future—and many are likely to be based on a two-sided model.

But two-sided markets are not just present in high technology; they are dotted throughout the economy. We began with perhaps a trivial example of dating clubs—discotheques, church clubs for singles, and local village matchmakers could have served just as well. Others range from real estate to video games to media firms. Some of the most recognizable brands in the world operate in two-sided markets: think of Bloomberg, Century 21, Sony and Nasdaq.

Two-sided firms behave in ways that seem surprising to those used to analyzing traditional industries, but in ways that seem like common sense once one understands the business problems they must solve. Firms must adopt price structures and investment strategies tailored to balancing the demands of the customer groups they must attract to their platforms – and then must induce to stay. That is a different (and harder) problem than those commonly faced by one-sided firms. American Express bet on a price structure skewed against merchants; it worked for many years, but eventually created great conflict. Visa has since surpassed American Express, a firm that was once dominant and seemed unbeatable.

Meanwhile, companies whose success we now take for granted made their mark by adopting price structures that originally seemed quite radical. Microsoft chose to cater to software developers. Bloomberg bet on a simple formula for its data terminals—a flat fee for subscribers and very modest charges for content providers.

There is no reason for regulators to steer clear of these industries or to scrutinize them with greater zeal. But they do need to be aware that different economic principles drive pricing and investment decisions in these industries. Prices do not—cannot—follow marginal costs in each side of the market. And price and investment strategies must optimize output by harvesting the indirect network effects available on both sides. Government failure to recognize these imperatives would put some of the most innovative firms operating in markets with exceptional productivity growth at risk.

INTERCHANGE FEES IN THE COURTS AND REGULATORY AUTHORITIES

Howard H. Chang,
NERA Economic Consulting

In the last two decades card association interchange fees have become a focus of scrutiny by the courts and competition and regulatory authorities on three continents. There has been agreement that some centrally determined interchange fee is necessary—the chaos that would result in the absence of an interchange fee has not been an appealing alternative. They have differed, however, as to whether the actual levels of interchange fees should be left to the card systems or whether the government should intervene.

In 1986, the U.S. Eleventh Circuit Court of Appeals ruled in *NaBanco* that Visa's interchange fee was procompetitive.[1] It agreed with a lower court, which had ruled that Visa's interchange fee permitted the system to operate more efficiently by eliminating costly negotiations among individual members and helped to solve imbalances between costs and revenues on the two sides of the system.

Fourteen years later, by contrast, the European Commission, acting on complaints made by associations of retailers, reached a preliminary determination that the interchange fee violated European laws against collective price setting.[2] On further reflection, the Commission decided there was no feasible alternative to a collectively determined interchange fee but it was unwilling to leave the determination in the hands of Visa. It reached an agreement with Visa requiring Visa to lower its interchange fee and to conduct cost studies, which would be reviewed by the Commission and would form the basis of cost-based benchmarks for the interchange fee.[3]

In 2002, the Reserve Bank of Australia (RBA)—which has certain regulatory authority over payment systems in Australia—also reached the conclusion that the interchange fee should be based on cost factors approved by government regulators, not privately determined by the card associations.[4] The RBA explicitly rejected economic arguments that it is important to account for demand factors in choosing relative prices in two-sided markets. The government bank is thus seeking to impose cost-based regulation of association interchange fees in Australia.

---

1 National Bancard Corp. v. Visa U.S.A., Inc., 596 F. Supp. 1231 (S.D. Fla. 1984), *aff'd*, 779 F.2d 592, 605 (11th Cir. 1986).

2 *Commission Plans to Clear Certain Visa Provisions, Challenges Others*, European Commission Press Release IP/00/1164, Oct. 16, 2000 (visited Nov. 19, 2002) <http://europa.eu.int/comm/competition/antitrust/cases/29373/studies/>.

3 *Commission Exempts Multilateral Interchange Fees for Cross-Border Visa Card Payments*, European Commission Press Release IP/02/1138, July 24, 2002 (visited Nov. 21, 2002) <http://europa.eu.int/comm/competition/antitrust/cases/index/by_nr_58.html>.

4 Reserve Bank of Australia, *Reform of Credit Card Schemes in Australia IV: Final Reforms and Regulation Impact Statement* (Aug. 2002), at 30-31 [hereinafter *RBA Final Reforms*].

Here, I look more closely at the role of interchange fees in two-sided payment card markets in the context of competition policy. Part A reviews the role of the "merchant discount"—the portion of a card charge not credited to a retailer's account when the charge is presented for collection—in the dynamics of competition among payment card systems. The absolute and relative levels of the merchant discount across systems have important implications for both competition and product differentiation.

Part B reviews the origin of Visa's interchange fee. This fee has been set collectively since the cooperative card systems were formed, in large part because the associations realized early on the fee was needed to balance the acquiring and issuing sides of the system.

Part C summarizes the *NaBanco* decision. The Eleventh Circuit recognized that collective price setting was necessary for the efficient provision of payment card services, just as collective price setting was necessary for the provision of the music royalty collection services provided by BMI.

Part D reviews the RBA's analysis of the role of interchange fees. Its analysis is wrong as a matter of economics, and has led to a policy recommendation that is likely to reduce consumer welfare. The discussion focuses on the RBA investigation rather than the European Commission investigation because there are more publicly available documents discussing the RBA's approach. The Commission's investigation was similarly flawed.

A. THE MERCHANT DISCOUNT, MULTIHOMING AND PAYMENT SYSTEM COMPETITION

The merchant discount determines merchant "demand" for card payment services—that is, the willingness of merchants to accept cards issued by a payment system. The merchant discount helps to position the payment card brand with merchants and cardholders. Thus, American Express has historically targeted high-end consumers and high-end merchants. It charged a merchant discount that was high relative to the card associations, but many high-end merchants were willing to pay this large discount because it enabled them to attract affluent customers and, perhaps more importantly, customers with corporate cards who were both big spenders and relatively insensitive to prices.

When Discover entered the payment card industry, however, its managers appealed to a broader group of consumers—indeed, as a subsidiary of Sears at the time, it was perceived by many as a low-end card. Discover chose to charge a merchant discount that was lower than either the associations or American Express. In part, this strategy was no doubt designed to build merchant acceptance quickly. But it may have also reflected the fact that Discover provided less value to merchants than the competition—most of its cardholders already had a card from another system and, as a group, their demographic characteristics did not make them more attractive to merchants than the alternatives.

The merchant discount is thus a weapon that the systems can use to alter their competitive positions, as well as a tool with which to respond to pricing changes by the other card systems and to deal with a variety of competitive issues that arise from what is called "multihoming." For example, as the payment card associations increased their acceptance rate among merchants and individuals during the 1970s and 1980s, American Express came under competitive pressure. Merchants and cardholders had multihomed—that is, most American Express accepting merchants also took Visa and MasterCard, and many American Express cardholders carried Visa and MasterCard. The resulting competitive tension became public in what came to be known as the "Boston Fee Party."[5]

Jasper White—then the owner of Jasper's on the Boston waterfront—led a group of Boston restaurateurs asking American Express to lower its merchant discount. American Express refused, and a few of the restaurants subsequently dropped the American Express card. The conflict generated national attention, and the other payment card systems used it as an opportunity to highlight their lower merchant discounts and larger card bases. American Express lowered its merchant discount dramatically in the following years.

The merchant discount has also been used by all the systems to manage entry into new segments. Most supermarkets refused to take payment cards through the 1980s, because the merchant discount cut too deeply into already low margins. The card systems eventually accommodated the supermarkets by offering them a lower merchant discount.[6] As three-party card systems, Discover and American Express did not need to coordinate the pricing policies of multiple card issuers and acquirers. They both directly lowered their merchant discounts to supermarkets. In contrast, the four-party systems, Visa and MasterCard, had to use interchange fees to accomplish the same end. Both Visa and MasterCard lowered their respective interchange fees for supermarkets; merchant banks ("acquirers") then lowered their merchant discounts to supermarkets. In four-party systems, the interchange fee is the instrument used to affect merchant discounts, as well as cardholder prices, that three-party proprietary systems set directly.

B. THE ORIGIN AND HISTORY OF VISA'S INTERCHANGE FEE[7]

Bank of America started its own credit card system in 1958, but banking regulations and other operational constraints limited its ability to expand the system beyond the borders of its home state, California. It thus chose to franchise

5 *See* DAVID S. EVANS & RICHARD SCHMALENSEE, PAYING WITH PLASTIC: THE DIGITAL REVOLUTION IN BUYING AND BORROWING 170-171 (1999).

6 *Id*. at 132.

7 This section is based in part on EVANS & SCHMALENSEE, *supra* note 5, at chs. 1, 3, 4 and 8.

the card brand, launching it nationally in 1966. Under the franchise system, Bank of America did not set the fees charged to cardholders and merchants by its licensees, so it, like the proprietary systems, needed an instrument to balance cardholder and merchant demand in its system. It chose to require the acquirer to turn over the entire merchant discount to the issuer on a transaction where the two banks were different. In other words, the implicit interchange fee under the Bank of America franchise system was equal to the full amount of the merchant discount. On a $100 transaction with a 5 percent merchant discount, for example, the merchant's bank would receive from the cardholder's bank only the $95 it had to turn over to its merchant. The issuer would keep the $5 merchant discount for itself.

Such an interchange fee clearly placed a relatively greater incentive on cardholder versus merchant acquisition. On a transaction involving different banks, the acquirer would receive no net revenues to cover its costs. However, given that most transactions in those days were "on-us" (the issuer was also the acquirer), a bank still had significant incentives to sign up merchants, as it would keep the entire merchant discount for "on-us" transactions.

Setting the interchange fee equal to the merchant discount might have struck the right balance for Bank of America's two-sided problem, but it posed significant operational difficulties because banks did not trust each other to report the full merchant discount charged. This interchange fee structure had "serious problems of uncertainty and instability…since it was based upon each merchant bank's interpretation of how much was due and owing to issuer banks."[8] There were significant additional problems beyond interchange. Financial losses, management problems, system inefficiency, and a distrust of Bank of America by the other banks, led to significant internal conflicts between Bank of America and its licensees. In 1970, Bank of America accepted a proposal from its licensees to transform the franchise system into a member-owned cooperative—National BankAmericard, Inc. (NBI), which subsequently changed its name to Visa. Using what amounts to an open membership policy, Visa has since built a global cooperative, with some 8,000 credit and debit issuer members in the United States alone.

To fix the interchange fee problem, Visa (then NBI) instituted a formal interchange fee that was uniform across members and was not rigidly linked to merchant discount fees charged by individual acquiring banks. The interchange fee was thus designed to perform a balancing function, "bringing the costs of the system in line with the revenues for each participating VISA member bank regardless of the role it plays, either merchant or issuer, in the VISA system."[9] The fee was initially based on costs but now depends significantly on demand factors as well.

8 National Bancard Corp. v. Visa U.S.A., Inc., 596 F. Supp. 1231, 1239 (S.D. Fla. 1984).

9 *Id.* at 1261.

The interchange fee is the only mechanism available to the card associations to determine the pricing structure collectively, and is the analog to the proprietary system's ability to directly set cardholder and merchant prices to balance demands. Individual association members determine all other prices for cardholders and merchants. Competition among these members tends to drive these prices down to marginal cost. Thus, the higher the interchange fee paid by acquiring banks to issuing banks, the greater the incentives for issuers to distribute cards and to lower prices to cardholders. By the same token, a lower interchange fee gives acquiring banks incentives to lower their merchant discounts, thereby increasing the number of merchants willing to accept the card. The interchange fee also provides an efficient mechanism for implementing association-wide marketing policies—for example, to increase acceptance by supermarkets or to increase the issuance of debit cards.

C. INTERCHANGE FEES ARE PROCOMPETITIVE: NABANCO V. VISA[10]

In a case filed in 1979, National Bancard Corporation (NaBanco) claimed that the interchange fee constituted a price fixing agreement that violated Section 1 of the Sherman Act. NaBanco specialized in signing up merchants and processing card transactions. When it acquired a card transaction from a merchant it was obliged by association rules to pay an interchange fee to the card-issuing bank. NaBanco noted that this was unlike check clearing—checks are exchanged "at par" as a result of the efforts of the Federal Reserve. It argued that the interchange fee was set by the Visa Board acting on behalf of issuing banks and was therefore illegal *per se*.

NaBanco also argued that it was illegal under a rule of reason because the interchange fee put banks that specialized in acquiring merchant accounts at an unfair disadvantage compared to banks that both serviced merchants and issued cards. It claimed that integrated issuer-acquirer banks could offer merchants a lower discount since these banks did not have to pay an interchange fee on transactions involving their own cardholders.

Visa did not deny that, in a literal sense, interchange fees established a price term among members. However, it relied on the BMI decision, in which the Court found that even though blanket royalty licenses were literally price-fixing, they were permissible under a rule of reason because of the efficiencies created.[11] Visa argued that, following BMI, the collective decision to set the interchange fee should be evaluated under the rule of reason. NaBanco rejected the application of BMI on the grounds that card association members had the practical option of negotiating interchange fees bilaterally.

---

10 This section is based in part on EVANS & SCHMALENSEE, *supra* note 5, at ch. 11.

11 *See* Broadcast Music, Inc. v. CBS, 441 U.S. 1 (1979).

Visa countered by explaining the role of the interchange fee in terms of what we would now call a two-sided market.[12] It portrayed the association as a joint venture of banks. The joint venture wanted to maximize the use of Visa by both cardholders and merchants. Visa argued that the purpose of the interchange fee was to provide a "mechanism to distribute and share the costs of the joint venture in relation to prospective benefits, thereby encouraging member to provide the Visa service to a competitively maximum extent on both the cardholder and merchant 'sides' of the business."[13] It also argued that the interchange fee was necessary for the joint venture to provide credit card services. The interchange fee was imposed to control opportunistic behavior by individual banks and to avoid a chaotic system involving literally thousands of bilateral negotiations between issuing and acquiring banks. Unlike classic price fixing, where ending collusion leads to higher output and lower prices, the outcome from eliminating interchange fees would be chaos, with lower output and possibly higher overall prices.

The district court and the Eleventh Circuit Court of Appeals agreed with Visa. The Eleventh Circuit upheld the district court's rule-of-reason approach in the case, explicitly relying on the two-sided nature of the industry:

> Another justification for evaluating the [interchange fee] under the rule of reason is because it is a potentially efficiency creating agreement among members of a joint enterprise. There are two possible sources of revenue in the VISA system: the cardholders and the merchants. As a practical matter, the card-issuing and merchant-signing members have a mutually dependent relationship. If the revenue produced by the cardholders is insufficient to cover the card-issuers' costs, the service will be cut back or eliminated. The result would be a decline in card use and a concomitant reduction in merchant-signing banks' revenues. In short, the cardholder cannot use his card unless the merchant accepts it and the merchant cannot accept the card unless the cardholder uses one. Hence, the [interchange fee] accompanies "the coordination of other productive or distributive efforts of the parties" that is "capable of increasing the integration's efficiency and no broader than required for that purpose."[14]

The Eleventh Circuit went on to find that "[a]n abundance of evidence was submitted from which the district court plausibly and logically could conclude that the [interchange fee] on balance is procompetitive because it was necessary to achieve stability and thus ensure the one element vital to the survival

12 One of Visa's consultants went on to write one of the earliest papers on the economics of two-sided markets. *See* William F. Baxter, *Bank Interchange of Transactional Paper: Legal and Economic Perspectives*, 26 J.L. & Econ. 541 (1983).

13 Brief of Appellee at 8 (citation omitted), National Bancard Corp. v. Visa U.S.A., Inc., 596 F. Supp. 1231, 1239 (S.D. Fla. 1984) (No. 84-5818).

14 National Bancard Corp. v. Visa U.S.A., Inc., 779 F.2d 592, 602 (11th Cir. 1986).

of the VISA system—universality of acceptance."[15] The Supreme Court declined to review the Eleventh Circuit's decision.

D. THE RESERVE BANK OF AUSTRALIA DEMURS

The RBA reached a different conclusion. It viewed Visa's interchange fee[16] as inherently suspect because the fee was set by horizontal agreement among competitors:

> [C]o-operative behaviour between competitors which involves the collective setting of prices is rarely permitted in market economies. Prima facie, such behaviour is anti-competitive and, where it is allowed, it typically requires some form of dispensation by competition authorities on the basis that there are offsetting benefits to the public.[17]

However, the RBA did not appear to seriously consider eliminating interchange fees, recognizing that "interchange fees can play a role in redressing imbalances between the costs and revenues of issuers and acquirers in four party credit card schemes."[18] Instead, it wanted to regulate because it was "not convinced that community welfare would be maximized if the setting of interchange fees…were left entirely to the schemes and their members in Australia[.]"[19] Specifically, it was concerned that Visa's interchange fees were set too high, thereby encouraging excessive use of credit cards as an alternative to other payment methods.[20] The RBA relied on the findings from theoretical models showing that firms may have private incentives to set interchange fees that are higher than the rate that maximizes the value of card transactions to society as a whole.[21] The models suggested that individual merchants have an incentive to accept Visa cards if they expect to make sufficient incremental sales, even if Visa cards are more expensive for the merchant than its alternatives.

The RBA argued that the private benefits to individual merchants associated with such incremental sales did not constitute a social benefit, as they came at the expense of sales by other merchants. For all merchants collectively, the

---

15 *Id*. at 605.

16 The interchange fees set by MasterCard and Bankcard, a domestic card association, were also at issue.

17 Reserve Bank of Australia, *Reform of Credit Card Schemes in Australia: A Consultation Document* (Dec. 2001), at 5 [hereinafter *RBA Report*].

18 *RBA Final Reforms*, *supra* note 4, at 30.

19 *Id*.

20 The RBA's investigation also concerned the card associations' rules that prohibited merchants from imposing surcharges for credit card use, as well as the associations' membership policies regarding entry.

21 *RBA Report*, *supra* note 17, at sec. 2.4.

reasoning goes, acceptance of Visa cards generates few if any incremental sales. Thus, collectively set interchange fees allow Visa to exploit each individual merchant's willingness to pay, which is derived from private rather than social benefits. The RBA thus proposed a regulatory scheme for interchange fees that was based on cost factors, rather than on demand factors. Moreover, it would consider only costs on the issuer side and would exclude many issuer side costs from consideration without providing an economic basis for such exclusion. Visa and MasterCard are challenging this in court.

There were three principal mistakes with the RBA's approach.[22]

First, the RBA failed to demonstrate a significant market failure resulting from interchange fees. Regulatory intervention should only be considered if such a showing can be made. The RBA did not establish that interchange fees exceeded the socially optimal level or that competition was somehow impaired. The RBA relied on theoretical models that showed the fees *could* be too high. But the same models also showed that privately set interchange fees *could* be at the socially optimal level—or *could* be too low. It makes no sense to seek to lower interchange fees when we do not know if they are too high today. All we do know for certain is that one side of the two-sided market—in this case, the merchants—would prefer to pay less.

The second problem with the RBA's proposed regulation is that it ignores a key implication of two-sided markets that its own expert, Michael Katz of the University of California (Berkeley), had acknowledged in his report:[23] there is no economic rationale for setting fees based solely on costs. As Katz noted, "there is little reason to believe that it is optimal to set the interchange fee equal to either an issuer's marginal costs of a card transaction or zero."[24]

In short, because the socially optimal interchange fee depends on both benefits and costs, regulation based on costs alone will not produce efficient pricing—except by chance. In fact, in evaluating various proposals for interchange fee setting, Katz cited a number of objections that apply directly to the RBA's proposal, which he did not analyze in his report. In particular, he criticized other proposals as attempts to "allocate costs based on functionality, with only vague reference to demand conditions."[25] The RBA plan would set a benchmark for interchange fees by allocating costs based on functionality, ignoring demand factors entirely—even though, as Katz explicitly stated, "efficient pricing must be based in part on demand conditions."[26]

22 The European Commission's approach was similarly flawed. As noted above, this section focuses on the RBA investigation because of the more detailed public record that is available.

23 Michael L. Katz, *Reform of Credit Card Schemes in Australia, II: Commissioned Report*, Reserve Bank of Australia (Aug. 2001), at 12-16.

24 *Id*. at 29.

25 *Id*. at 34.

26 *Id*. at 35.

The third problem with the RBA's approach is its exemption of proprietary systems, such as American Express and Diners Club, from the proposed regulatory scheme:

> American Express and Diners Club, on the other hand, do not have collectively determined interchange fees. Whether they have an internal transfer mechanism or "implicit" interchange fee is not relevant; the three party card schemes do not have a process under which competitors collectively agree to set a price which then affects, in a uniform way, the prices each of the competitors charges to third parties.[27]

This reasoning reflects a lack of understanding of two-sided markets, and in the process creates an unjustifiable competitive disadvantage for the open card associations. As noted earlier, Visa and MasterCard use the interchange fee to balance costs and demands on the two sides in the same way that proprietary systems achieve balance by directly setting prices to end-users. If it is not anticompetitive for American Express to use two-sided pricing, it should not be anticompetitive for Visa to do the same.

The fact that Visa's decision is a collective act by its members, who are horizontal competitors, does not imply that it has chosen a pricing structure that harms the collective interests of its customers. Indeed, if the concern is that merchants pay too much compared to cardholders (and thus encourage overuse of the card), the matter should be of even greater concern for the American Express system: American Express merchant discounts are generally steeper than Visa discounts—by about a third in the United States.[28]

The RBA did not explain why the horizontal nature of Visa's association structure made the use of two-sided pricing any more problematic than the same strategy used by proprietary systems. It thus failed to recognize the importance of the interchange fee to the card associations in competing in a two-sided industry.

E. CONCLUSIONS

It is hardly surprising that retailers would prefer to pay lower discounts on card payments, or that they have attempted to enlist regulators and courts to fight for lower fees on their behalf. Nor is it surprising that governments have taken their complaints seriously. For one thing, merchant trade associations pack considerable political clout.

---

27 *RBA Report*, *supra* note 17, at 118. In fact, as the RBA noted, AMP Bank also issues American Express cards in Australia. *Id.* The RBA did not explain why American Express was exempted from the proposed regulatory scheme.

28 United States v. Visa, 163 F. Supp. 2d 322, 333 (2001).

But a close look at the mechanism suggests that, as courts in the United States have recognized, the interchange fee is critical to balancing demand in a two-sided market. Unlike classic price-fixing, the collectively determined fee does not generally raise prices or lower output. It reflects the same demand and cost factors as the socially optimal interchange fee. In fact, the two coincide in non-trivial cases. On the other hand, the alternative proposals—cost-based regulation, the imposition of a zero fee, or a ban on interchange fees—are demonstrably sub-optimal.

INTERCHANGE FEES: A REVIEW OF THE LITERATURE

Richard Schmalensee,
MIT Sloan School of Management

The literature on the economics of interchange fees has developed rapidly in recent years. Two forces have stimulated these writings. Controversies surrounding these fees have led a number of economists and lawyers—sometimes acting as experts for the parties involved—to study how and why these fees are set and what the consequences are for consumers. At the same time, economists have begun to recognize that two-sided markets have fascinating and hitherto unexplored economic characteristics, and that the analysis of interchange fees can provide insights into these markets generally. This paper summarizes the key contributions to the relevant literature and is limited to papers that have been published in professional journals (or are forthcoming at the time of this writing).

A. BAXTER (1983)[1]

In a paper derived from his work as an economic expert in *NaBanco*, William Baxter, a Professor of Law at Stanford University who was then serving as head of the Justice Department's Antitrust Division, performed what seems to be the first economic analysis of interchange. Baxter provided a model of the supply and demand for payments system services, along with analysis of the way interchange fees evolved in the context of check-based and credit card-based transactions.

Baxter postulates a four-party, two-sided market consisting of (1) a merchant who receives "transactional paper" as payment for goods, (2) the merchant's bank, which deposits the payment to the merchant's account, (3) the purchaser of the good who presented the transactional paper to the merchant, and (4) the purchaser's bank, which "contemplates acceptance of and payment against" the purchaser's paper. Note that the transactional "paper" in this model could as easily be an electronic transfer as a personal check or a credit card purchase slip.

The merchant deposits the purchaser's liability in his bank account. The merchant's bank, in turn, demands payment from the purchaser's bank, which debits the purchaser's account or requires the purchaser to transfer funds in some other way. The services required to get all this done have real costs, including the time of bank personnel and the carrying cost of the computer and telecommunications infrastructure employed. Hence, one can think of a cost-related supply function for bank transaction services.

1 William F. Baxter, *Bank Interchange of Transactional Paper: Legal and Economic Perspectives*, 26 J.L. & ECON. 541 (1983).

Similarly, the service yields benefits for both the merchant who accepts the payment and the purchaser who makes the payment. Some of these benefits are straightforward—e.g., eliminating the need for both the merchant and the purchaser to hold cash. And some may depend on the nature of the contract governing the transaction—e.g., who bears the risk of default and who gains interest-free use of the float.

Baxter noted, however, that a critical feature making this market distinctive is its two-sided nature. The transactional service for any particular transaction is consumed by both the purchaser and the merchant, and they must together cover its marginal cost. Similarly, it is supplied jointly by the two banks involved, and they must each receive enough to cover their individual marginal cost. Neither the purchaser nor the merchant will pay more than what a particular service is worth to her alone. Furthermore, the service won't be delivered if their combined offer of payment is less than the two banks' combined marginal cost of performing it.

Baxter stressed that in market equilibrium, the same number of transactions (indeed the exact same transactions) must be agreed to by purchasers, merchants, merchants' banks, and purchasers' banks. He also pointed out that if purchasers pay their banks and merchants pay their banks, there is no guarantee that this balance will be attained. Consider a simple example involving a single transaction. Suppose that the marginal cost for each bank to execute the transaction would be 2¢, that the value of executing the transaction to the merchant would be 5¢, and that the value of the transaction to the purchaser would be 1¢. Even though executing this transaction would produce a net social benefit of 2¢ [=(1+5)-(2+2)], it will not be executed if the purchaser's bank only receives revenue from the purchaser, since the purchaser is only willing to pay 1¢. The problem is easily solved in this case. The merchant's bank could charge the merchant 3¢ and remit 1¢ to the purchaser's bank, which would charge the purchaser 1¢. Each bank's costs are covered, the transaction is thus executed, and the merchant retains a net benefit of 2¢. Other arrangements will also do the trick in this case, but they all involve a payment from one bank to another—an interchange fee.

In Baxter's model, it is generally necessary to use an interchange fee to shift payments between purchasers' and merchants' banks in order to balance the system. "To describe the activities traditionally performed by one bank or another," Baxter writes, "is not to say that the costs of these activities must be borne by the bank performing them." Thus, typically, Baxter suggests, "there must be some particular side payment between merchant bank and purchaser bank…that will bring the receipts of each bank into equality with the marginal cost it has incurred…"

Baxter is careful not to specify the direction of this interchange payment; that, he understood, depended on both supply and demand factors. Baxter also was aware that the price of the marginal transaction, the division of the burden

between merchant and purchaser, and the division of the receipts between merchant and purchaser bank all turn on market conditions.

Baxter complements this theoretical analysis with two case histories of "four-party transaction vehicles." The first describes the development of check and similar "draft" clearing mechanisms in the United States over the past century and a half, focusing on the question of why the incidence of interchange fees changed and why interchange fees subsequently fell to zero. Or, to put it in the language of banking, how the practice of clearing checks "at par" came to be institutionalized.

Before the Civil War, check-like drafts were used largely to pay merchants in distant cities and typically involved payment of large interchange fees by the purchaser's bank. After the passage of the National Bank Act in 1864, however, there was a rapid shift to interchange fees paid by merchants' banks (and thus, indirectly by merchants). Baxter links this shift to two broad factors affecting the four parties: a sharp fall in transportation and communication costs that reduced the total cost of clearing checks and the growth of scale economies in clearing that differentially favored purchasers' banks.

Later, the rise of private clearinghouses serving many banks both reduced clearing costs and increased incentives to standardize interchange fees that in earlier eras had been established through bargaining. Baxter contends that low, standardized interchange fees would have been entirely compatible with efficiency. However, when the Federal Reserve consolidated the clearinghouse function, it conditioned use of the mechanism on accepting checks at par. And since the Federal Reserve system was both very efficient and indirectly subsidized by the government, most banks found it in their interest to go along. Nonetheless, there were still 1,547 "non-par" banks operating in 1964, and they only disappeared altogether in 1980.

The second case history follows the use of interchange fees in the evolution of payment cards from the 1950s to the 1980s. The early cards, Baxter explains, were "travel and entertainment" cards targeted at high-income consumers, oil company cards usable only at affiliated gasoline retailers, and bank cards accepted by a variety of merchants—but only within the bank's legal deposit-taking region. Fees charged to merchants and cardholders varied enormously, and in some cases were very high by contemporary standards. But since these were all "three-party" cards—the owner of the card system serviced both the merchant and the cardholder—there was no need for the owners of the transactions services to make side payments in order to balance the two sides of the market.

Four-party payment card transactions only arose after 1966, when the Bank of America licensed its BankAmericard system nationwide, and other banks were authorized to service merchants and/or purchasers. In 1966, a group of banks organized a cooperative to perform network services for a card system that eventually became known as MasterCard. And in 1970 the BankAmericard

system was reorganized along parallel lines to form what would later be named Visa.

Baxter notes that in any such system, once a servicing bank pays the merchant for a purchase, it is at the mercy of the purchaser's bank unless it has a contract defining the terms at which it will be compensated. Thus a bank "cannot be permitted to announce daily the price at which it will buy paper to be billed to cardholders." Merchants' banks and card-issuing banks might, in theory, solve this problem with fees negotiated bilaterally in advance of transactions. But Baxter points out that with even a modest number of banks, the number of bilateral agreements would be unmanageably large. With just a dozen banks, for instance, 132 bilateral agreements would need to be negotiated and implemented.

Moreover, he notes that there would be a free-rider problem that encourages opportunistic behavior. It would benefit individual card-issuing (i.e., consumers') banks to charge more than the optimal system-wide fee in the expectation that a single bank's higher charge would have relatively little impact on the average interchange fee and thus little impact on merchants' willingness to accept all cards bearing the association's brand. The problem of estimating optimal fees for cards is complicated, he adds, by the reality that some cardholders make use of the line of credit attached to the cards—and are thus more valuable customers for the issuing bank—and some don't.

"The courts," Baxter concludes, "should recognize that collective institutional determination of the interchange fee is both appropriate and desirable." And while "this collective process of equilibration resembles horizontal price-fixing, it should not be so treated" because "individual establishment of interchange fees will almost certainly produce chaotic results, such as higher fees and instability within the card systems."

B. CARLTON AND FRANKEL (1995)[2]; EVANS AND SCHMALENSEE (1995)[3]

Baxter's views on the consequences of payment card interchange fees and the rules under which they were determined were challenged by Dennis Carlton of the University of Chicago and Alan Frankel of Lexecon, an economic consulting firm. Their analysis, part of a broader exploration of the role of antitrust in regulating the payment card associations, was written in the context of *NaBanco*, mentioned above.

2 Dennis W. Carlton & Alan S. Frankel, *The Antitrust Economics of Credit Card Networks*, 63 ANTITRUST L.J. 643 (1995); and Dennis W. Carlton & Alan S. Frankel, *The Antitrust Economics of Credit Card Networks: Reply to Evans and Schmalensee*, 63 ANTITRUST L.J. 903 (1995).

3 David Evans & Richard Schmalensee, *Economic Aspects of Payment Card Systems and Antitrust Policy Toward Joint Ventures*, 63 ANTITRUST L.J. 861 (1995).

NaBanco, an agent for merchants' banks in both the Visa and MasterCard associations, sued to stop Visa members from setting the interchange fee collectively, arguing that the single fee gave an unfair advantage to banks that served both merchant and card-issuing sides of a transaction. But the courts, following Baxter's reasoning, found that "Visa established that [the interchange fee] is necessary to offer the Visa card—a procompetitive benefit which offsets any anticompetitive effects."

Carlton and Frankel argue that Baxter's analysis doesn't go far enough: under perfectly competitive conditions, without frictions, "interchange fees will have absolutely no effect on ultimate prices or the ability to compensate the issuing bank for any costs." In this ideal case, consumers who use credit cards may be asked to pay a surcharge or be given a discount, and this can serve to cover the costs of the system. To see how this works, recall the simple example above. Suppose the merchant's bank charges the merchant 2¢, and the merchant offers the purchaser a 1¢ discount to have the system in question execute the transaction. Then if the purchaser's bank charges 2¢, her net cost is 1¢ (after deducting the 1¢ discount), exactly as before, the merchant pays 3¢ (including the 1¢ it pays to the purchaser), and each bank receives 2¢ and thus just covers its costs.

Carlton and Frankel then claim that interchange fees may be harmful under imperfectly competitive conditions. Card-issuing banks may not be forced to compete away revenues in excess of costs that are generated by interchange fees, for instance. They may keep the money, raising the net price of using cards and restricting output of the card services industry. Or they may spend the money on excessive promotion of cards, increasing the use of cards at the expense of more efficient payment mechanisms.

Even if competition among issuing banks is intense, so that these problems can be ruled out, they note that merchants may be prevented by card association rules from placing an efficient surcharge on card use—one that is equal to the merchants' net cost of accepting cards in payment—or may simply choose not to do so to avoid extra costs of posting multiple prices. In this case, the interchange fee must be recovered through an increase in the average price of all goods, regardless of how they are purchased. This increase acts like a tax on the use of cash or other alternatives to cards and distorts incentives to use various payment mechanisms. They seem to suggest that the interchange fee should be set to zero to avoid this, though they do not explicitly recommend this.

Writing in a subsequent issue of the *Antitrust Law Journal*, David Evans of NERA Economic Consulting and Richard Schmalensee of MIT take issue with Carlton and Frankel's analysis of the implications of frictions and market imperfections.

Merchant discounts for using cash would be economically equivalent to surcharges for using credit card, Evans and Schmalensee note, and cash discounts would not violate card association rules. Yet such discounts are quite rare, they point out, most likely because the bookkeeping costs and loss of goodwill outweigh the potential benefits. Under these conditions, the size of the interchange fee does indeed have an impact on who bears the cost of operating the card payments system and thus, in general, affects the system's total output. Evans and Schmalensee argue, however, that there is no reason to believe that an interchange fee of zero is closer to the optimum than the positive fees agreed upon by the card associations.

Evans and Schmalensee also address the issue of the cross-subsidy between cash and card customers implied by no-cash-discount merchant policies. They acknowledge that this may, indeed, distort incentives in the choice of payment mechanisms. But retail markets are full of small distortions of this sort. Because offering cash discounts would entail real costs and in the presence of many similar distortions, it is far from clear that the net impact on balance of no-cash-discount policies is reduced efficiency and consumer welfare. Indeed, singling out this distortion, as opposed to dozens of others associated with retailers' decisions to limit the unbundling of services from sales, is peculiar.

They also point out that the economic effects of no-cash-discount policies depend on the merchant discount, not the interchange fee. Thus American Express, which, as a single corporate entity, has no need of an interchange fee, charges a higher merchant discount than Visa or MasterCard. To force reduction or elimination of interchange fees is to distort competition and choice between these different sorts of payment systems.

Carlton and Frankel then respond to Evans and Schmalensee, again in the *Antitrust Law Journal*. They note that much of the criticism from Evans and Schmalensee concerned interventionist positions Carlton and Frankel did not believe they had stated and, in any event, did not hold. In particular, they agree that doing away with the interchange fee and going to bilateral negotiations would be a bad idea. They point out that while they suggested the potential for anticompetitive effects from interchange fees, they had not conducted the type of detailed study that would be needed to conclude that overall anticompetitive effects existed. Without such evidence, they do not believe that intervention would be appropriate. However, they criticize Evans and Schmalensee for concluding that interchange fees were not anticompetitive without having done a detailed study themselves.

C. FRANKEL (1998)[4]; CHANG AND EVANS (2000)[5]

Like Baxter, Alan Frankel looks to banking history for evidence of the welfare consequences of interchange fees. But in emphasizing different parts of the tale, he draws very different conclusions.

Baxter largely attributes the disappearance of interchange fees for checks to a combination of regulatory carrots (subsidies in the maintenance of a government-run clearinghouse), regulatory sticks (pressure from the Federal Reserve to use the zero-interchange government clearinghouse), and falling transactions costs (improved transportation and communications sharply reduced the real resources needed to clear a check). Frankel, by contrast, sees it as a consequence of Federal Reserve pressure to abandon cartel pricing of the services provided by private local clearinghouses. The "at par" collection of checks is efficient in spite of the costs of the clearing process, he says, because individual banks can and do charge customers (competitive) fees for the service.

Interchange fees for checking survived longest in isolated, one-bank towns, he argues, because these banks had the most market power. Not only did they not face local competition in clearing checks, they were able to free-ride on distant banks in competitive local markets that were reluctant to pass on the occasional interchange fee to a checking customer.

Frankel emphasizes the impracticality of merchants putting surcharges on card transactions or offering discounts for cash because of a combination of legal restrictions, contractual restrictions imposed by card associations, and high costs in maintaining multiple pricing schemes. This strong tendency to "price coherence," he says, creates market power for the card associations in setting interchange fees. Card users are insensitive to the magnitude of the fees because they pay the same price as cash customers. While merchants don't share their indifference, they may be willing to pay more to the banks than the net resource saving associated with card use, because card acceptance attracts profitable customers.

The prospect that card-issuing banks may well compete away the resulting economic rents by offering rebates, promotional considerations and the like to attract customers gives Frankel only modest comfort. He stresses that even if competition among merchants is perfect, those who pay with cards get services for less than their cost, while those who pay cash pay more than cost, leading to overuse of cards and under-use of cash, and perhaps, other competing payment methods.

---

4 Alan S. Frankel, *Monopoly and Competition in the Supply and Exchange of Money*, 66 ANTITRUST L.J. 313 (1998).

5 Howard H. Chang & David S. Evans, *The Competitive Effects of the Collective Setting of Interchange Fees by Payment Card Systems*, 45(3) ANTITRUST BULL. 641 (2000).

To avoid this distortion, Frankel would ideally require merchants' and customers' banks to execute transactions "at par"—with a zero interchange fee.

Howard Chang and David Evans of NERA Economic Consulting take issue with criticism of private, collective interchange fee setting found in Carlton and Frankel (1995) and Frankel (1998). They begin with a description of how four-party payment card transactions are executed, explaining how the actions of association members must interact to achieve both traditional economies of scale and the network economies associated with balancing the merchant and cardholder sides of the two-sided market. Network effects lead to a chicken-and-egg problem: a brand new system is effectively worthless to merchants because there are no cardholders to use cards and worthless to cardholders because there are no merchants who will accept cards. Once the system is off the ground, coordination of the two sides of the market, fine-tuned with interchange fees, maximizes total value. And alternatives to private, collective rate-setting of interchange fees—bilaterally-negotiated fees, a ban on fees that forces issuing banks to recoup all their costs from cardholders, government-regulated cost-based fees—all have serious drawbacks in terms of generating excessive transactions costs, failing to internalize external benefits and costs, and distorting incentives.

They stress several points of general relevance. First, fees charged to merchants by acquirers in the card associations are lower than fees set by American Express, the largest proprietary system. If, as Frankel asserts, collective determination of the interchange fee generated market power, that power should be reflected in higher merchant discounts. Yet the average merchant discount for Visa, which has a card base that is six times larger than American Express, is about one third lower than that of American Express. Moreover, Visa's interchange and merchant discount fees fell during the first few decades of the association's rapid growth.

Second, regulation of card association interchange would put the associations at an artificial disadvantage with respect to proprietary card systems. The bottom line for both Carlton and Frankel (1995) and Frankel (1998) is the desirability of placing limits on interchange fees. Yet, interchange fees are not needed—indeed, can play no part—in setting charges for merchants and cardholders in closed, three-party proprietary systems such as American Express and Discover. Thus, any restrictions on interchange fees (let alone a ban) would bias regulation in favor of the closed systems. This is at least ironic in the case of American Express, which, because of its higher merchant discounts, must in Frankel's view have done more than the bank associations, dollar for dollar, to undermine efficient pricing of transactions services.[6]

6 Discover, the other large proprietary system, has generally charged lower merchant discounts than the associations. Nonetheless, on average its discounts have exceeded the markups charged by bankcard acquirers over the associations' interchange fees. Thus if Visa and MasterCard were forced to eliminate interchange fees and Discover were not regulated, Discover's merchant discounts could well end up higher on average than those of the bank associations.

Third, they argue that analogies between card payment interchange and check interchange are misleading. They note that par clearance of checks came about only through substantial government intervention—banks could gain access to the Federal Reserve's national clearing system, at subsidized fees, and other benefits only if they agreed to par clearance. They also note that there is no analysis or evidence to support the proposition that the Federal Reserve's success in driving interchange rates to zero represents a competitive, welfare-maximizing equilibrium in check clearing. For example, check authorization (providing some verification or guarantee of available funds) and check truncation (making a check into an electronic transaction) that are desired by some consumers and merchants might have developed more quickly with different interchange fees for checks.

D. SCHMALENSEE (2002)[7]

The papers discussed thus far either assume perfect competition among banks or allude to imperfections in competition without formally modeling them. Richard Schmalensee presents an explicit model of imperfect banking competition, in which a bank cooperative sets the interchange fee in order to maximize a weighted sum of the profits of card-issuing and merchant-servicing banks with some market power. After the cooperative has acted, the banks set prices to consumers and merchants in order to maximize their individual profits.

In order to focus on the role of interchange as a balancing instrument in payment systems, he does not model retailer competition or consider the implications of the "price coherence" argument. He assumes that the two sets of banks face declining demand curves for their services and that these demand curves (which are assumed to be linear in most of the analysis) can be connected to social welfare as in standard analyses of markets such as those for sugar or electricity. As discussed below, Rochet and Tirole (2002) and Wright (2001) relax this assumption by imposing explicit models of consumer behavior and of bank and merchant competition.

In Schmalensee's model, total system volume depends on the product of the number of merchants accepting cards and the number of consumers who carry and wish to use cards. Thus, if the merchant discount is very high, for instance, few merchants accept cards, and the value of the system to consumers and to card-issuing banks will be low even if many consumers carry cards. Similarly, if charges to consumers are set very high, the system will be worth little to merchants and the banks that serve them even if many merchants want to accept cards. By, in effect, shifting costs from one side of the system to the other, the interchange fee enables the cooperative to steer between these poles to

7 Richard Schmalensee, *Payment Systems and Interchange Fees*, L(2) J. INDUS. ECON. 103 (2002).

enhance its members' profits. The central question addressed is whether this sort of unregulated collective price determination should be thought of as cartel behavior with a thin coat of varnish or whether, as Baxter argued, "collective determination of the interchange fee is both appropriate and desirable."

The results of this analysis generally support Baxter. While the main effect of cartel behavior is to harm consumers by restricting output, in a special case (but without any extreme assumptions) of the Schmalensee model, collective interchange fee determination maximizes output and social welfare in order to maximize the system's private value to its owners. While this does not occur in all cases, as a general matter both the privately and socially optimal interchange fees are determined mainly by *differences* between the demand, cost, and competitive conditions faced by card-issuing banks and those faced by merchant-servicing banks. Banks' markups are determined by the competitive conditions they face; the optimal use of the interchange fee is mainly to increase volume to the benefit of all parties.

The interchange fee that maximizes private value may be above or below the fee that maximizes total system output, and if the value-maximizing fee is above (below) the output-maximizing fee, so is the welfare-maximizing fee. Schmalensee shows that in general, no cost-based approach to regulating interchange fees is guaranteed, even in theory, to enhance social welfare. This analysis reveals no economic case for requiring the interchange fee be set to zero or for prohibiting the use of any interchange fee.

A key insight from the Schmalensee model is the distinction between viewing payment services as an upstream input in a vertical market in which merchants are the customers, and a two-sided market in which both merchants and goods buyers consume payment card services. This latter way of looking a payment cards indicates how the collective setting of interchange fees increases market efficiency.

E. ROCHET AND TIROLE (2002)[8], (2003)[9]

The work of Jean-Charles Rochet and Jean Tirole of the Institut D'Economie Industrielle in Toulouse, France is the first to model explicitly the behavior of all actors in a four-party payment system. This structural approach in Rochet-Tirole (2002) requires fairly strong simplifying assumptions for tractability but permits a fully rigorous analysis of bank, consumer and merchant behavior and of the determinants of the relation between market equilibrium and social welfare.

8 Jean-Charles Rochet & Jean Tirole, *Cooperation Among Competitors: Some Economics of Payment Card Associations*, 33(4) RAND J. ECON. 549 (Winter 2002).

9 Jean-Charles Rochet & Jean Tirole, *Platform Competition in Two-Sided Markets*, J. EUR. ECON. ASS'N (forthcoming 2003).

First, Rochet and Tirole assume that banks serving merchants are perfectly competitive while, as in the Schmalensee model, card-issuing banks are each assumed to have some market power. Second, customers are assumed to make a fixed number of purchases. Consumers choose whether or not to carry a card, where to shop, and, if they choose a store that accepts cards, whether to pay with a card or the alternative payment system (which I will generally call "cash" for convenience). Third, it is assumed that the fees charged by issuers decrease with the interchange fee, so that at least some of the higher revenue is competed away in better terms for cardholders. This assumption is consistent with a wide range of models of bank competition, even a single monopoly issuer. Fourth, consumers are assumed to differ in the benefit to them of paying with the card rather than cash. It follows that the number of cards carried falls as cardholder fees rise. Fifth, a specific, standard model of retailer competition is assumed.

In the first stage in this model, the interchange fee is set—either by the profit-maximizing association or by a welfare-maximizing regulator. In the second, issuers set card fees and consumers decide whether to hold a card. At the same time, merchants decide whether to accept cards under terms offered by the banks and set prices for their own products. In the third stage, consumers observe merchants' prices and whether cards are accepted, then choose a store and whether to pay with a card or with cash.

Rochet and Tirole concentrate on the case in which all merchants are identical. In this case, the card association will charge the highest interchange fee that keeps all merchants on board. (Since the business of servicing merchants is assumed to be perfectly competitive with constant costs, only issuing banks earn profits and thus only they care about the interchange fee.) Rochet and Tirole then demonstrate that if merchants cannot (for whatever reason) offer cash discounts and if the alternative payment system is provided and priced efficiently, this interchange fee is either socially optimal or leads to an over-provision of credit card services. Over-provision can arise when competition among merchants provides strong incentives to accept cards in order to capture business that would otherwise have gone to a competitor, even if marginal consumer benefits are only weakly affected. This enables issuing banks to charge a high interchange fee without losing merchants, and the proceeds from the high fee are used, at least in part (because of competition), to set inefficiently low prices to consumers.

It is important to note that even with cash discounts ruled out and with rent-seeking competition among merchants inflating the incentive to accept cards, over-provision is not inevitable, contrary to the assertions of Carlton and Frankel and Frankel. Rochet and Tirole go on to make clear that no-discount rules by themselves are never sufficient for over-provision. They do this by assuming that some fraction of consumers is uninformed as to which stores accept cards. In deciding whether or not to accept cards, stores recognize that their choice will not affect the shopping behavior of this uninformed segment,

and the business-stealing incentive to accept cards is thereby reduced. Rochet and Tirole then show that even if over-provision occurs for some set of parameter values when all consumers are informed, if enough consumers were instead uninformed, the privately optimal interchange fee would also be socially optimal. They also show that increased competition among issuing banks tends to make over-provision more likely by leading to reductions in consumer fees.

If the association chooses an interchange fee that is too high, it is of course possible *in theory* to improve matters by regulation. However, the Rochet-Tirole analysis makes clear how difficult it would be to attempt to improve matters in practice. Even in their simplified model, there is no guarantee that setting the interchange fee equal to zero or basing it on the costs of issuing and acquiring banks would produce a gain in social welfare. In the simple numerical example above, for instance, the fact that both issuing and acquiring banks have costs of 2¢ would likely persuade most regulators that the interchange fee should be zero, even though at an interchange fee of zero the system in that example is not viable. Even in the simplified Rochet-Tirole model with only cash and a single card system, socially optimal interchange fees depend on benefits to consumers and merchants that are difficult to measure, as well as on the exact nature and intensity of competition among issuing banks and among merchants. Any serious attempt to improve on association-determined interchange fees in the real world would have to employ more complex models and to confront even more daunting measurement problems.

Rochet and Tirole provide a preliminary analysis of a model with two card associations in addition to cash that illustrates this last point. Because a merchant who declines to accept the card of only one association will not lose the business of all consumers who carry at least one card, the incentive to accept cards is less inflated by merchant competition in this case. Nonetheless, Rochet and Tirole show that competition between cards need not result in a lower interchange fee, and, if it does, this reduction may lower social welfare. They also show that when two cards compete, allowing merchants to impose discounts or surcharges (and assuming away frictions that may nonetheless prevent them from doing so) may increase or decrease social welfare. It is clear that modeling the real world, in which two card associations compete with cash and checks and, in the United States, two proprietary systems and a variety of debit cards, would reveal additional levels of complexity, in the face of which regulatory determination of interchange fees or the removal of no-surcharge rules would improve social welfare only by purest chance.

In a recent paper that provides important new insights into two-sided markets, Rochet and Tirole (2003) generalize the multiplicative demand model introduced by Schmalensee (2002) and study the outcome of competition between two credit card associations (and more generally between two platforms in two-sided markets) They compare the resulting price structure (in particular, the allocation of costs between the two sides of the market) with the social

optimum. Again, the determining factors are the same (demand and competitive conditions on both sides of the markets involved, along with costs) but the competitive price structure generally differs from the optimal one. Importantly, though, and consistent with the earlier literature, there is no systematic bias. For example, in the case of linear demands (this was also true under some conditions in Schmalensee (2002) and Wright (2001)) the two price structures coincide.

Rochet and Tirole also analyze the factors that determine the direction of interchange fees (or, more generally, which side of the market pays a greater share of revenues) and apply it to a series of mini case studies. Based on this preliminary review, they find that the case studies provide some encouraging support for the theoretical framework.

F. WRIGHT (2001)[10], (2003)[11]

Julian Wright of the University of Auckland in New Zealand extends the Rochet-Tirole framework. Wright (2001) drops the assumption of identical merchants by assuming a continuum of industries, across which merchants differ in the benefits received from accepting cards but within which all merchants are identical. Wright also applies the Rochet-Tirole treatment of imperfect competition among issuing banks to both issuing and acquiring banks, generalizing the approach of Schmalensee. Like Rochet and Tirole, Wright derives the behavior of all actors in the system from first principles. Because merchants are not identical, increases in the interchange fee reduce the number of merchants who accept cards, all else equal. For the most interesting portion of his analysis, Wright assumes the same form of competition among merchants as Rochet and Tirole.

The timing in Wright's model is similar to that of Rochet-Tirole. After the interchange fee is set, issuers and acquirers set prices, and then merchants decide whether to accept cards and consumers decide whether to use them. In equilibrium, within each industry all merchants will either accept or refuse cards. Similarly, a consumer who decides to use the card will use it for all purchases in all industries that accept it.

Wright initially considers the general determinants of the interchange fees that maximize (a) output of card payment services, (b) total profits of payment card association members, and (c) social welfare. In Rochet-Tirole, the second

10 Julian Wright, *The Determinants of Optimal Interchange Fees in Payment Systems*, UNIVERSITY OF AUCKLAND DEPARTMENT OF ECONOMICS WORKING PAPER #220 (2001). This paper came after those by Schmalensee and Rochet-Tirole, but the working paper version is dated earlier. It is under submission to the JOURNAL OF INDUSTRIAL ECONOMICS.

11 Julian Wright, *Optimal Card Payment Systems*, EUR. ECON. REV. (forthcoming 2003).

of these was always greater than or equal to the third. Here, as in the Schmalensee model, increases in the interchange fee can reduce profit by lowering merchant acceptance, and any ordering of these three fees seems possible, in general.

Many of Wright's most interesting results rest on assumptions that imply linear merchant and consumer demands for card services. Under those assumptions, when customers don't know if merchants accept cards, accepting cards does not attract others' customers, and if competitive conditions are the same for issuers and acquirers, the output-maximizing, profit-maximizing, and welfare-maximizing interchange fees are identical. If consumers are better informed, merchants have stronger incentives to accept cards, and (at least when competitive conditions are the same for issuers and acquirers) it is profit-maximizing to take advantage of this by charging a higher interchange fee. Interestingly, in at least one class of cases, the welfare-maximizing interchange fee is also higher. The socially optimal response to over-acceptance by merchants for business-stealing reasons in this model is to *raise* the interchange fee to discourage over-adoption.

Like Schmalensee and Rochet-Tirole, Wright finds that profit-maximization may produce the socially optimal interchange fee. He thus illustrates, again, the profound difference between collective determination of the interchange fee and ordinary cartel price determination. Moreover, as in the earlier papers, when the profit-maximizing fee is not welfare-maximizing, the difference between them depends on a host of factors that would be difficult, at best, for any regulator to measure and integrate.

In Wright (2003), he takes the Rochet-Tirole model and considers the extremes of monopoly merchants and perfectly competitive merchants. Among other things, Wright shows how these assumptions constrain the ability of card schemes to use interchange fees or no-surcharge rules in ways that harm social welfare. He argues that since these extremes (very concentrated or very competitive merchant markets) may be the most relevant to when surcharging will actually arise, they demonstrate the positive role of the no-surcharge rule.

G. BALTO (2000)[12]; AHLBORN, CHANG AND EVANS (2001)[13]

David Balto, a partner in the law firm White and Case and a former head of the policy office of the Federal Trade Commission's Bureau of Competition,

12 David A. Balto, *The Problem of Interchange Fees: Costs without Benefits?* 4 EUR. COMPETITION L. REV. 215 (2000).

13 Christian Ahlborn, et al., *The Problem of Interchange Fee Analysis: Case without a Cause*? 22 EUR. COMPETITION L. REV. 304 (2001).

examines the current relevance of the rationale for permitting collective setting of interchange fees found in the 1984 *NaBanco* case.

In that case, the court rejected NaBanco's argument that Visa's interchange fee amounted to illegal price-fixing for several reasons: (a) the fee was needed to recover costs that might not otherwise be recoverable, (b) the accounting evidence suggested the fees were cost based, (c) no less-collusive method of determining fees was practical, (d) competition between payments modes limited fee-setters' market power, and (e) fees were largely internal transfer payments, since individual banks were both merchant acquirers and car issuers. Balto finds fault with each of these.

(a) The need to recover costs: Balto notes that payment card clearance costs have fallen sharply since the early 1980s. Moreover, electronic processing has also made it more practical for issuing banks to charge cardholders directly as an alternative to recovering costs indirectly through merchants.

(b) Reliance on accounting evidence of a cost basis for fees: Balto is both skeptical of the premise that antitrust officials are able to oversee cost-based pricing and wary of the incentives created by what amounts to cost-plus price regulation. Moreover, he notes that interchange fees have risen even as costs have fallen.

(c) The lack of a less restrictive alternative: Bilateral negotiation of fees between acquiring and issuing banks has become more practical, Balto says, because both sides of the market are much more concentrated than they were in the 1980s. What's more, since he says that issuing banks are capable of collecting fees directly from cardholders, it follows that a system with no interchange fees at all would now be possible. And a no-interchange standard would eliminate the need for bilateral negotiations.

(d) Limits on payment card market power: Visa and MasterCard have much more market power than seemed likely in the 1980s, Balto says. New network entry is now difficult, he adds, and merchants are very reluctant to lose the strategic advantages of giving their customers the option of paying by card. The fact that Visa could increase interchange fees on debit card transactions by 10 percent without losing significant business, he says, proves the point.

(e) Interchange fees as a "neutral transfer payment:" Balto notes that banks have largely withdrawn from the merchant side of the transactions, so the associations now have incentives to favor card issuers.

Balto does not specify his preferred alternative to the institutional rules sanctioned by *NaBanco*. However, he suggests that a no-interchange rule, with banks collecting costs directly from customers, would be relatively efficient. He entertains other possibilities, as well: (1) allowing interchange fees as a means to solve the chicken-egg problem during a limited start-up period for networks, (2) regulating fees according to costs, (3) limiting cost-justified fees

to narrow, well-defined cost categories, (4) limiting debit card interchange to per-transaction fees, (5) making clear that banks are free to bypass association interchange rates and set their own bilaterally, (6) eliminating all merchant non-discrimination rules—including no-surcharge rules—so that customers have appropriate incentives to use alternate payment modes.

Christian Ahlborn (Linklaters), Howard Chang and David Evans respond to the Balto arguments in a subsequent issue of the *European Competition Law Review*. The core of their argument is that Balto has overlooked the forest in weighing the impact of the trees—that interchange fees serve a vital function as a mechanism for internalizing the externalities of a two-sided market.

Any payment card system, be it a three-party proprietary system or a four-party bank cooperative, confronts similar issues in balancing fees paid by merchants and cardholders. Any system will thus have good reasons to price discriminate (in the economist's benign usage), charging more to users of the service with lower elasticities of demand. Any system will also take into account network effects—the need to use differential fees to balance participation by merchants and customers. And, of course, fee-setting in any system will be influenced by both the magnitude of costs and the division of costs in servicing merchants and cardholders.

The interchange fee is an instrument for allocating total system cost and revenue between merchants and cardholders. But, the authors note, it "is not a price paid by the acquirers (and thus indirectly by merchants) for services rendered by the issuers." This idea, they point out, is based on the misconception that payment card services fit into a standard vertical market structure in which "upstream" issuers supply inputs to "midstream" acquiring banks, which in turn provide services to "downstream" merchants. In fact, cardholders are consumers of payments services, too. The market is two-sided, and the interchange fee "accounts for the relative importance of merchants and cardholders in developing the system."

This view is supported by historical evidence. Visa and MasterCard built their huge heterogeneous merchant networks by offering much lower merchant discounts than American Express and Diners Club. And they induced merchants to purchase terminals for servicing online debit transactions by charging lower fees for those transactions.

Ahlborn, Chang and Evans also challenge the idea that it has become practical to substitute bilaterally-negotiated interchange fees. While credit card issuing and merchant acquisition activities may be more concentrated than they were in the 1980s, they contend that there are still far too many parties involved to make this practical. In any event, individual banks still have no incentive to take into account costs and benefits that are external to them, and, as Baxter pointed out, individual issuers have a strong temptation to free ride by inflating their interchange fee demands. These are critical issues in markets with network economies.

By the same token, a zero-interchange-fee rule imposed by regulators would leave the card associations without an instrument to balance the two sides of the market—something that is available to all firms in two-sided industries (such as real estate, video games, computer operating systems, and exchanges). Indeed, they argue, it would greatly favor three-party proprietary systems, which would have the merchant discount available for balancing cardholder and merchant demand for their systems. And it might lead large banks to abandon the four-party systems, undermining the economies of scale that originally led to their formation.

More generally, the authors point out that there is no reason to equate collectively-set interchange fees to horizontal price fixing in traditional markets, which is anticompetitive on its face. They note that even if competition among banks is imperfect, banks will not be able to translate interchange fees directly into excess profits: competition will force issuers to spend revenues that exceed servicing costs on promotion of cards and/or benefits to cardholders in the form of rebates and ancillary services.

More directly, if setting interchange fees was merely cartel behavior, it would surely never lead to *over*-provision of payment system services, though the Rochet-Tirole analysis shows that over-provision can occur even in a relatively simple model. Moreover, Ahlborn, Chang and Evans argue, interchange fees give bank associations a chance to compete on even footing with three-party proprietary systems like American Express, which don't need transfer payments to balance the two sides of the market or to exploit economies of scale.

In Balto's view, the fact that few merchants offer discounts for cash implies that the market for payment modes is distorted in favor of cards. However, Ahlborn, Chang and Evans argue that merchants' failure to price goods in bundles is just one of myriad forms of minor market failure—think of the distortion of "free" parking in malls. In any event, the only practical fix here, they say, would be a zero-fee rule, which would undermine efficiency to a greater extent by undermining network economies.

H. CONCLUSIONS

The economic literature on interchange fees is very young and deals with a complex and unusual phenomenon. The older literature tries to address it without rigorous modeling. Thus, it was possible even after Baxter's seminal contribution to suggest, as did Carlton and Frankel, that society might be better off mandating a zero interchange fee. The more recent literature still leaves many theoretical and empirical questions unanswered. However, several conclusions can be drawn from the efforts to model interchange fees formally by Rochet and Tirole, Schmalensee and Wright.

1. While there is no guarantee that interchange fees set by associations will maximize social welfare, the cost and demand factors driving private fee-setting are closely related to those that determine socially optimal fees. Moreover, any deviation from social optimality will be the result of subtle differences between the two sides of the market—not from market power leading to excess profits for association members. No rigorous analysis supports the idea that a zero interchange fee or a fee set by regulators based on cost would generally raise welfare. In all the rigorous analysis in the literature, socially optimal fees depend on demand and competitive conditions, as well as on costs. Nor is there any support for the notion that bilateral rate-setting by banks, even if practical, would be likely to produce an equilibrium closer to the social optimum.

2. The fact that collective, private interchange fee determination cannot be depended upon to generate exactly the optimal levels of payments services doesn't imply that collective rate-setting is appropriate grist for antitrust scrutiny. The hallmark of anticompetitive pricing is excess profits and less-than-optimal output. Yet, there is no reason to believe that private setting of interchange fees generates excess profits for the "colluding" parties since competition in acquiring and issuing activities can be expected to dissipate any surplus revenue. The interchange fee has the effect of lowering costs on one side of the market and raising them on the other. Nor will interchange fees set above the socially optimal level generally restrict output. Indeed, critics of private rate-setting argue that output tends to exceed the optimal level because cardholders pay less than the real cost of the transactions they initiate and merchants are willing to pay more than the savings associated with card payments in order to capture business from each other.

3. No-surcharge rules leading to cross-subsidies from cash to card customers may distort the relative output of card and cash payments. But this is inevitable in the "second-best" world in which we live, in which other incentives are also distorted. Nor is there reason to believe that, in the absence of no-surcharge rules, merchants would "unbundle" the net increase in transactions costs associated with card payment services and charge card customers accordingly. Indeed, merchants have always found it in their interests to bundle services in a variety of ways—for example, stores may include "free" parking and no-charge shipping with products, restaurants include bread and drinks refills with meals, etc. There is no evidence that the distortion resulting from "same price, cash or charge" is substantial compared to distortions caused by other common service-bundling practices.

BANK INTERCHANGE OF TRANSACTIONAL PAPER: LEGAL AND ECONOMIC PERSPECTIVES

William F. Baxter,
Stanford University

Previously published in Journal of Law and Economics, Vol. 26, No. 3, 1983.

Abstract

Consumer purchases by means other than currency—for example, by check, credit card, or debit card—generate a paper record that must be handled by the merchant, the merchant's bank, the purchaser's bank, and the purchaser. Before coming to Washington, I was involved in several controversies over the terms on which these types of records would be created and exchanged between banks. That involvement led me to think that economics provides novel and useful insights into the process of interchange and the payment systems of which they are a part.

In this article I examine some of those lessons. I focus primarily on the economics of financial institutions in generating and exchanging accounting information essential to the operation of four-party cashless payment systems. Section I develops the economic theory of these systems, and Section II examines the evolution of four-party cashless payment systems in the light of this theory.

---

* Assistant Attorney General, Antitrust Division, United States Department of Justice. This paper was written while I was Professor of Law at Stanford University and revised thereafter. The views expressed here are my own and are not official policy statements of the Antitrust Division or the Justice Department. I thank J. Anthony Chavez and Greg Sidak for their helpful research assistance and suggestions.

[Journal of Law & Economics, vol. XXVI (October 1983)]
© 1983 by The University of Chicago. All rights reserved.
0022-2186/83/2603-0004$01.50

I. THE THEORETICAL VIEWPOINT

The payment systems I discuss all involve four parties and four consensual arrangements. For example, in the checking context, the parties are the payee of the check, the bank in which the payee deposits the check for credit to his account, the bank on which the check is drawn (typically a bank with which the maker of the check has a depository arrangement), and finally, the maker of the check, usually a depositor with the drawee bank. In the context of the credit card or the debit card, four functionally analogous parties are involved, although the labels attached to them differ.

Because I focus on what is common to these payment mechanisms rather than on the distinctions between them, I use neutral terms to describe the actors and operations inherent in these mechanisms—terms not associated with any particular payment mechanism. Each payment system generates certain accounting information, which is exchanged among the four parties in order to facilitate an exchange of goods or services between two of the parties. (Although electronic signals soon may replace much of the paper that embodies the accounting information required for cashless payment systems, this would not affect the basic economic issues addressed in this article.) For convenience, I refer to the embodiment of this accounting information as *transactional paper* regardless of its physical form, and to the generation and exchange of transactional paper as *transactional services*. I assume that the person who initially receives the transactional paper is a *merchant* (*M*) who receives it in payment for goods; I refer to the bank in which he deposits the paper for credit to his account as the *merchant's bank* (*M* bank);[1] I assume that the person who gives the paper does so in his capacity as *purchaser* (*P*) of the goods sold by the merchant; and I refer to the bank with whom the purchaser has an arrangement that contemplates acceptance of and payment against that paper as the *purchaser's bank* (*P* bank). Nothing turns on the assumption that the purchaser and the merchant are in fact playing those particular roles. What is critical to the analysis is that there are at least four parties and that their relationship to the payment mechanism is analogous to the one I have described.[2]

A. The Demand for Transactional Paper

Any bargained-for exchange requires *P* to pay *M* for goods or services received. Once an economy moves beyond barter, the concept of payment involves much abstraction. Even if *P* tenders the gold coins of the realm, *M* is willing

1 Like "transactional paper," for the purpose of this article "bank" is an abstraction for financial intermediaries. It includes savings and loan associations that process "NOW account" paper and credit unions that process "draft account" paper.

2 I say at least four parties because often additional banks or clearing houses participate in the process, facilitating the flow of the transactional paper from the merchant's bank to the purchaser's bank. For the most part, whether additional parties participate is irrelevant to the basic points.

to accept the coins not because *M* can use them to fashion jewelry or fill his teeth but because he expects other merchants to "honor" the coins—that is, to be willing to deliver goods and services which *M* wants in exchange for the coins. The progression from gold coins to bank notes, to negotiable paper, to credit card charge slips, to electronic impulses as acceptable forms of payment makes clear that what is involved is a mechanism for causing multiple accounting entries to be made in several different sets of books, entries that in their totality constitute the community's recognition of each person's entitlements to consume. Merchant *M*, having delivered goods to *P* at an agreed price, wishes to have his consumption credits enhanced on the books of the community by the amount of the price; and since the rules of the community require that books balance, *P* agrees to have the consumption credits posted to his name reduced by an equal amount. Adjustments of the community's books in crediting *M*'s account and in debiting *P*'s account on the occasion of a purchase are accounting services that facilitate the needs of both the merchant and the purchaser. In terms of supply and demand, *M* and *P* have demands for transactional services in order to effect the appropriate entities in the community's books; banks supply such services.

Although a given transactional service may have as its fundamental purpose adjustment of the accounts of *M* and *P*, it will also have a variety of other product characteristics, such as cost of supply, convenience to the consumer of service (whether *M* or *P*), speed of adjustment, and accuracy of entry. There is no prior reason to believe that the preferences of merchants for a given transactional service would be the same as that of purchasers or even that different merchants (or purchasers) would have identical preferences. Consequently, the distribution of transactional services in terms of their product characteristics, the prices for these services, and the volume of their production are all questions remaining to be answered in the context of a market equilibrium.

At first impression transactional services appear to be private, not public, goods. Banks are able to extend such services to those who are willing to pay for them, whether merchants or purchasers, and to exclude from the services those who are not. Yet transactional services are unlike most private goods, because one cannot determine the aggregate (or industry) demand for them in the traditional way by horizontally summing the individual consumers demands.

Demand for a private good depends on each person's evaluation of the good's marginal utility and can be described by a function indicating the amount of product the person is willing to buy at a given price. Each consumer's evaluation of the marginal utility of a private good is usually independent of other consumers' evaluations, and so aggregate demand at any price level is the sum of the individual demands at that price. For example, if the prevailing price of shoes is $30 a pair, consumer Jones will buy one, and then another, and then another pair of shoes until the marginal value he attaches to the next pair (which he does not buy) falls below $30. The same is true for consumer Smith,

although there is no reason to expect that at any particular price each will demand the same number of pairs, because there is no particular reason to suppose that the marginal value that Jones attaches to the third or fifth or eighth pair of shoes is the same as the marginal value that Smith attaches. Because the evaluations of the marginal value of shoes by Jones and Smith are independent of one another, the aggregate demand of Jones and Smith for shoes at $30 a pair is simply the sum of their individual demands at that price.

In the case of transactional services, however, although consumer *P's* marginal valuation of the additional use of a particular payment mechanism may differ markedly from consumer *M's* marginal valuation,[3] these valuations cannot be independent of one another as in the case for shoes. The mechanics of transactional services require that for every transaction in which a purchaser becomes a maker of a check, there must be one— and precisely one—transaction in which a merchant becomes a payee; similarly, each use of a credit card by a card holder must be matched by precisely one act of acceptance of the card (or, more accurately, the paper that the card generates) by a merchant.

This identity in the type of transactional service used by the merchant and purchaser in a given exchange introduces a constraint not normally found in markets for private goods and reflects the interdependence in the marginal valuations between merchants and purchasers. Because the mechanics of transactional services require the acceptance of a particular payment mechanism by *both* the merchant and the purchaser to effect any given purchase, the marginal valuation of a transactional service by one party to the purchase is contingent on the acceptability of this form of service by the other party. On the one hand, given that particular payment mechanism is acceptable to the other party, marginal valuation is determined in the usual manner for private goods. On the other hand, if the payment mechanism in question is unacceptable to the other party for whatever reason, the marginal valuation by the first party is zero regardless of the magnitude of its value when the mechanism is acceptable. The contingent nature of these marginal valuations of transactional services by merchants and purchasers, and hence the contingent nature of the individual demands for these services, destroys the independence necessary to permit the calculation of aggregate demand by summing the individual demands horizontally and largely renders intractable the economics of transactional paper in this particular description of the market.

Perhaps the most intuitively appealing way to resolve the difficulties posed by this market model is to redefine what we mean as one unit of the product consumed. Rather than considering the demands of *P* and *M* as demands for separate products, define one unit of product to consist of the bundle of

3 Note that although *P* and *M* have a consumer-supplier relationship with respect to one another, they are both *consumers* with respect to transactional services, which in my nomenclature are supplied by banks.

transactional services that banks must supply jointly to *P* and *M* in order to facilitate the execution of one exchange of goods or services between *P* and *M*. Under this interpretation, the supply price of the product is the sum of the individual charges to *P* and to *M*. Furthermore, the demand for that product is a joint demand of *P* and of *M*: in combination they must make a payment of that magnitude to the banks to induce the necessary supply, but independently neither *P* nor *M* necessarily confronts any particular price as one he must pay in order to have his demand fulfilled.[4] This model preserves the excludability property of transactional services.

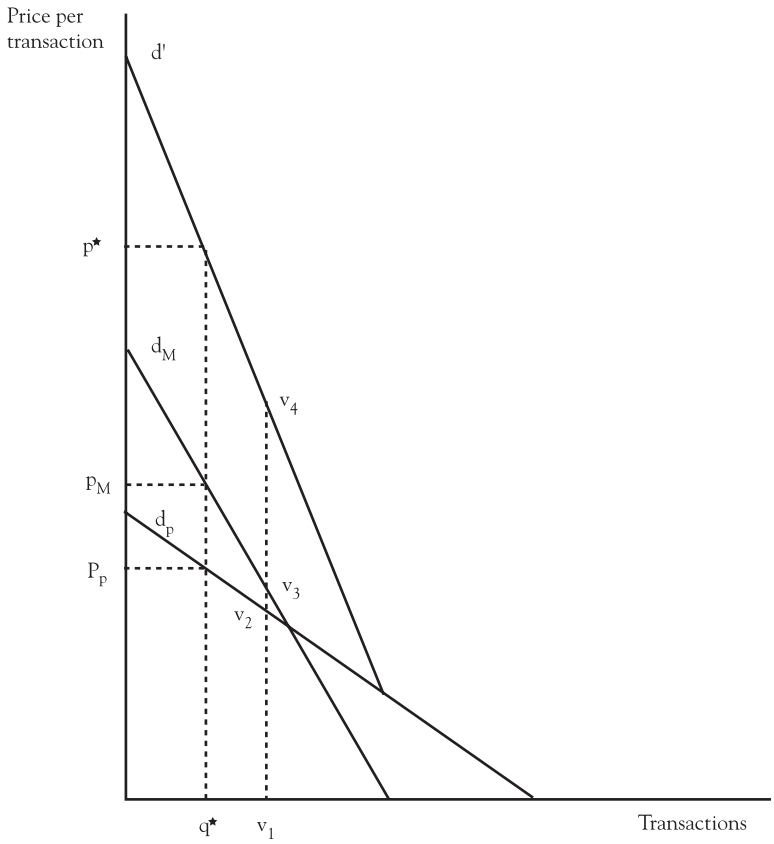

FIGURE 1

4 Another way of viewing the problem is to consider the transactional services provided to *P* and those provided to *M* as separate products that are jointly consumed, analogously to joint consumption of public goods. It is now widely recognized that the analytical apparatus long used in dealing with joint-cost problems also has application to peak-load pricing problems and to public good problems. The critical common feature is that the demand schedules of consumers must be summed vertically rather than horizontally in order to derive aggregate demand. This technique can be traced in the literature at least as far back as Howard R. Bowen, The Interpretation of Voting in the Allocation of Economic Resources, 58 Q. J. ECON. 27 (1943).

Figure 1 illustrates the derivation of aggregate demand for transactional services of a given type in a single-merchant, single-purchaser economy. The quantity axis is calibrated in units which represent the bundle of services that must be provided by banks to both *P* and *M* in order to facilitate one exchange. The vertical axis gives the reservation prices of the two traders for various levels of consumption of the transactional services. Line d_M represents the demand schedule of *M* for such complete units of transactional service on the assumption that *P*—*M's* customer—is willing to use this particular service but unwilling to make any contributory payment for the units when purchased from the bank. Line d_P represents the demand schedule of *P*, based on the assumption that *M* is unwilling to make any contributory payment for those services. Given the information shown in line d_M and line d_P, the aggregate demand schedule of *M* and *P* for these units of transactional services is line d', which is obtained by summing vertically the separate demand schedules of *M* and *P*. In other words, the schedule d' is constructed so that if any vertical line is drawn through the figure, the distance v_1v_4 equals the sum of distances v_1v_2 and v_1v_3.

Figure 1 should be interpreted as follows: if the price per complete transaction—that is, the total revenue banks will demand to provide the services necessary to facilitate one exchange between *M* and *P*—is p^*, then the quantity of transactions that *M* and *P* should demand is q^*, the quantity indicated by a vertical line dropped from the intersection of p^* and d'. I say "should" rather than "will" be demanded because, although q^* is the quantity of transactions that maximizes the aggregate benefits of *M* and *P*, a certain amount of coordination is prerequisite to *M* and *P's* arriving at that outcome. Specifically, this favorable outcome will result only if the aggregate price p^* is apportioned between *M* and *P* in the proportions represented by the height of their respective demand curves at output level q^*. That is, for each transaction, *P* must find a way to make some payment p_P to the banks, and *M* must find a way to make some payment p_M to the banks; when p_P and p_M are summed they will, by construction in Figure 1, equal p^*, the price that the banks demand for providing those services. If there are no bargaining costs—that is, if *P* and *M* have perfect information and neither persists in strategic bluffing to reduce his own costs at the expense of the other—they would bargain to this particular outcome. On the other hand, if either *P* or *M* strategically insists on paying less, then, because the other can be induced to pay no more at so high a level of transaction services, both *P* and *M* will be harmed, for the sum of their contributions will be less than p^*; thus the banks will decline to provide services that *M* and *P* together value at p^*.

One must resist any impulse to say that *M* is paying too much and *P* too little in the circumstances depicted by Figure 1. Given that the banks will insist on receiving revenues per transaction in the amount p^*, and given that *P* is unwilling to pay more than p_P per transaction at output level q^* for the very good reason that he does not value the service any more highly, *M* can only worsen his position by declining to make a payment per transaction in the amount p_M. For it is inescapable that *M* and *P* must agree on some specific

number of transactions to be effected by the payment mechanism in question. And if that number is to be q^*, then in our hypothetical case depicted in Figure 1 agreement can only be reached if *M* is willing to pay the preponderant share of the price p^*. In the region q^*, *M* values the marginal transaction more highly than does *P,* and *M* pays accordingly.

In our example, the individual demand schedules imply that if the level of transaction prices required by banks fell substantially, *M's* valuation of these transaction services would decline more rapidly than would *P's*. There is a particular output level, corresponding to the intersection of the individual demand curves where equal contribution would be required for equilibrium. And there is a still higher output level at which *M* would be unwilling to pay anything for additional services: to the right of that point *P* would have to bear all bank-imposed charges in order for equilibrium to be attained.

Figure 1 depicts how the individual demand schedules of a particular merchant and purchaser must be aggregated vertically in order to obtain a well-defined expression of the aggregate demand for transaction services in this miniature economy. However, since in our model merchants trade only with purchasers and not with other merchants, as we increase the number of merchants beyond one we must sum their individual demand schedules *horizontally* to obtain the aggregate merchant demand schedule. Similarly, if more than one purchaser exists in the economy, we must sum their individual demand schedules horizontally to obtain the aggregate purchaser demand schedule. Then, as in our one-merchant, one-purchaser case, the total aggregate demand schedule in the multi-merchant, multi-purchaser economy is obtained by summing *vertically* the two partial aggregate demand schedules of the two classes of traders.

The multi-merchant, multi-purchaser case is illustrated in Figure 2. Although the total number of transactions demanded industry-wide will be orders of magnitude larger than that depicted in Figure 1, Figure 2 retains the basic feature of Figure 1: merchant demand and purchaser demand are each depicted individually, and the aggregate demand for transaction services that confronts all participating banks in the community consists of the vertical aggregation of these two partial aggregate demands. For it remains true in the industry context, as in the case of the individual merchant, that a transaction is a two-sided arrangement, that transaction services facilitate the needs of both merchant and purchaser, and that agreement on a common number of transactions to be effected through the particular payment mechanism will not be possible with an equal division of charges between merchants and purchasers except under the extremely unlikely coincidence that the aggregate level of charges per transaction required by the banks lies directly above the intersection of those separate demand curves.[5]

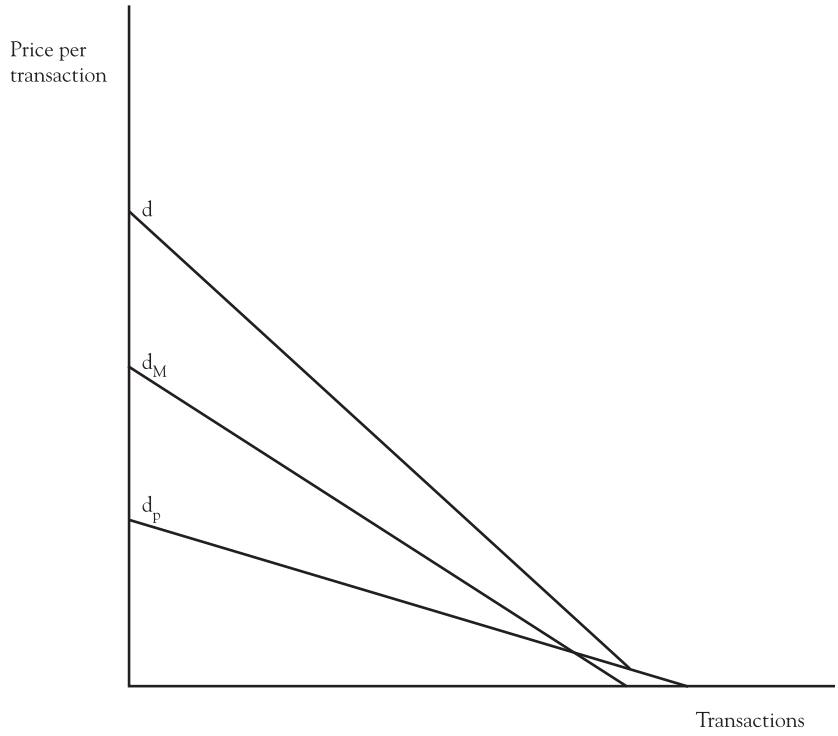

FIGURE 2

B. The Supply of Transactional Paper

A polarity corresponding to that of *M* and *P* on the demand side exists on the supply side as well: *P* has his banking relationship with one institution, *P* bank, and *M* has his banking relationship with another, *M* bank.[6] Both *M* and *P* bank will incur costs associated with establishing the payment system and providing services essential to effecting each transaction between *P* and *M*.

5 Indeed, in any real-world setting there may be no such intersection, although in my diagrams I have drawn the separate curves so as to produce one. It is not unlikely that in the real world the demand curve of merchants lies everywhere above, or perhaps everywhere below, the demand curve of purchasers, in which case there is no possible equilibrium that entails an equal division of transaction costs.

6 The assumption that there are precisely two banks adopted to facilitate discussion. In actuality there will be some number of purchaser-merchant transactions in which both parties to the transaction happen to have their banking relationships with the same financial institution. Some of the problems discussed in this paper arise in that context. There will be other transactions in which more, perhaps many more, than two banks will be involved—for example, when transactional paper is forwarded through a series of correspondent relationships for ultimate clearance. While these cases present additional problems, substantially all of the analytically difficult problems that arise on the supply side are present in the two-bank situation. Accordingly, I ignore the possibility of multibank clearance chains.

One can identify a set of activities that, at least in the typical case, will be performed by the employees of *M* bank, in principal part at *M*'s business premises. Such activities include soliciting, negotiating, and executing contractual agreements with merchants who do business in the geographical vicinity of *M* bank; participating in the periodic delivery by merchants to *M* bank of *M*'s records of transactions with purchasers; entering on the books of *M* bank credits to the account of M; capturing, in one form or another, the identity of the purchasers with whom *M* dealt and the identity of *P* bank with whom each *P* has his banking relationship; forwarding those data through some interchange or clearance mechanism to *P* bank; and bearing the cost of capital to the extent that unconditional credits are posted to *M*'s account before payment is received from *P* bank.

Analogously, there will be certain activities that typically will be performed by the employees of *P* bank, in major part at its business premises: soliciting, negotiating, and executing agreements with purchasers who wish to use the payment mechanism; receiving from a large number of *M* banks data about transactions executed by those purchasers; posting debits to the individual accounts of its various purchasers; transmitting periodic statements of those accounts to its various purchasers; and, in the case of arrangements not involving antecedent deposits by purchasers, receiving payment from those purchasers and entering credits to their account corresponding to their payments; bearing the costs of capital to the extent that unconditional credits are forwarded to *M* banks before payment from purchasers is in hand; and bearing the risk of purchaser default.

To describe the activities traditionally performed by one bank or another is not to say that the costs of these activities must be borne by the bank performing them. Just as it is true on the demand side that there must be an identity between individual purchaser transactions and individual merchant transactions, so also is it true on the supply side that there must be an identity between individual merchant bank transactions processed and individual purchaser bank transactions processed. For example, signing up merchants would be pointless if purchasers were not simultaneously being signed up. Hence, on the supply side, the costs of the activities of *M* bank and *P* bank must be regarded as joint costs with respect to each individual transaction, in the same sense that, on the demand side, demand of merchants and purchasers is strictly interdependent.

Correspondingly, the geometry of aggregate supply is analogous to that of aggregate demand. It is conventional to think of the supply curve for an industry as being constituted by the horizontal aggregation of the supply curves of the individual firms. But because the costs incurred by the banks are joint, when *P* bank participates on behalf of purchasers and *M* bank participates on behalf of merchants, the costs of the two firms must be aggregated vertically, not horizontally, in order to obtain an analytically useful representation of the

full marginal cost per transaction and hence of the number of purchaser-merchant exchanges that banks will facilitate at any particular price level for transactional services.

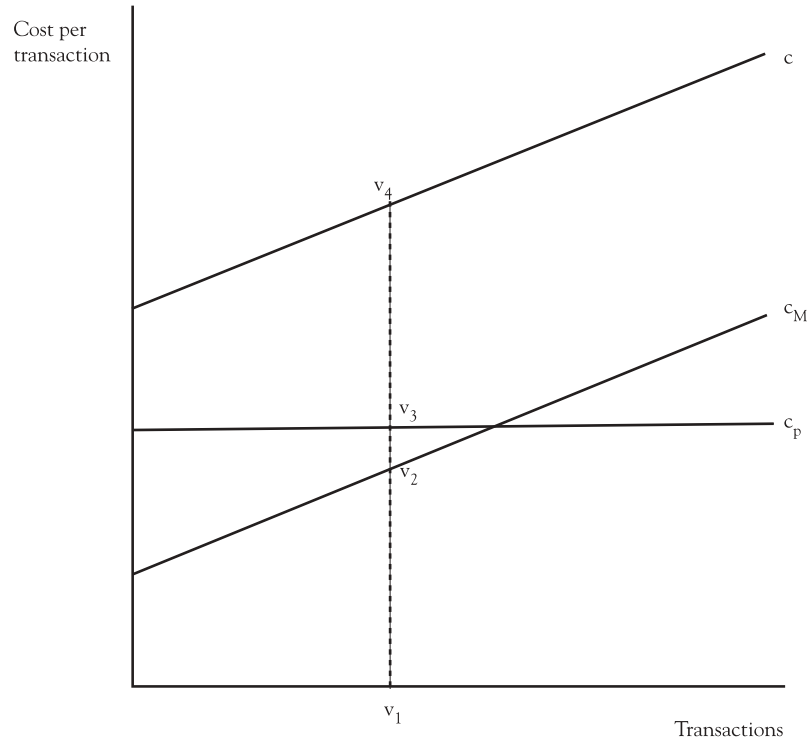

FIGURE 3

Figure 3 depicts possible marginal cost curves c_M for *M* bank, and c_P for *P* bank, together with their vertical aggregation *c*, which corresponds to the total marginal cost per exchange facilitated by the two participating banks. As before, the technique of vertical aggregation is such that, given any vertical line drawn through the curves, the distance v_1v_4 equals the sum of the distances v_1v_2 + v_1v_3. Somewhat arbitrarily, I have drawn Figure 3 in a way that suggests that *P* bank's costs exhibit constant returns to scale whereas *M* bank's costs exhibit decreasing returns to scale, but nothing in the analysis turns on those particular assumptions.[7] Figure 3 also could be thought of as depicting industry supply, if one views c_P as a traditional horizontal summation of the marginal

7 The analysis would be significantly affected if *C* exhibited negative slope over a very wide range. That would be the result if both c_M and c_P, had negative slope over that range or if either c_M or c_P had negative slope over that range to a degree that exceeded the positive slope of the other. If *c* had negative slope through the range of equilibrium output, the existence of natural monopoly conditions would be strongly suggested.

cost curves of all purchaser banks, and c_M as the traditional horizontal summation of marginal cost curves of all merchant banks. But in this interpretation, too, the vertical summation c of those two sets of costs depicts the industry supply curve, for with respect to each transaction, revenue equal to c must be forthcoming in order to cover all industry marginal costs.

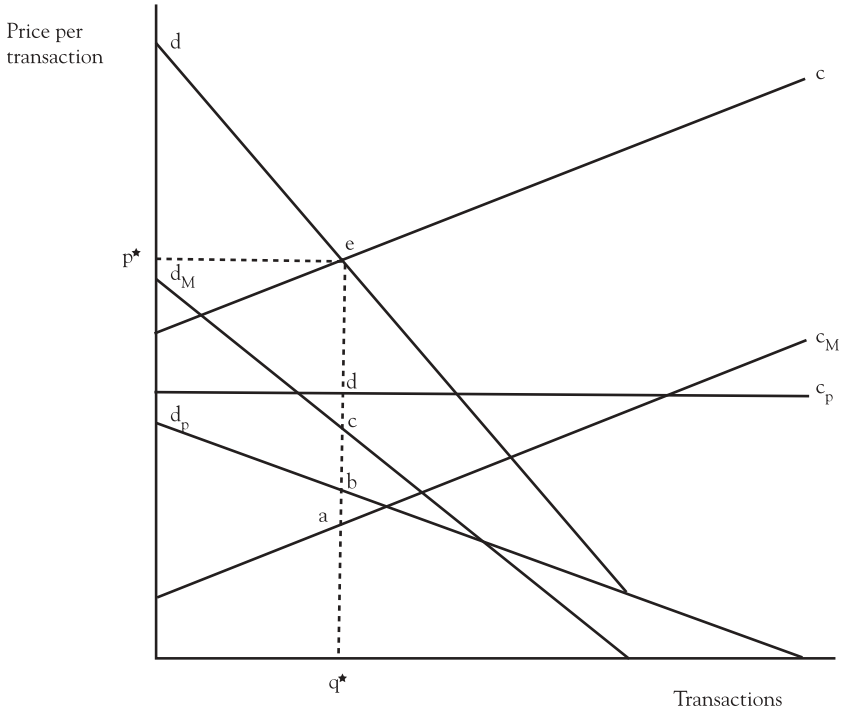

FIGURE 4

FIGURE 4. — Merchant makes sales of amount S; M bank discounts q^*c; merchant gets $S - q^*c$; P bank collects $S + q^*b$ from purchaser; together banks retain $(S + q^*b)_P + (-S + q^*c)_M = q^*b + q^*c = q^*e$; P bank remits $S + q^*b - q^*d$ to M bank. At close,

P's position	$-S$	$-q^*b$			
P bank position	$+S$	$+q^*b$	$-S$	$-q^*b$	$+q^*d$
M bank position	$-S$	$-q^*c$	$\pm S$	$+q^*b$	$-q^*d$
M's bank position	$+S$	$-q^*c$			
Totals down	0	0	0	0	0

Totals across:

P bank	$+ q^*d$ = cost
M bank	$q^*c + q^*b - q^*d = q^*a$ = cost
"Interchange fee"	$(q^*d - q^*b) = (q^*c - q^*a)$

Figure 4 depicts the resulting demand-supply equilibrium. In view of the total marginal cost per completed transaction, the industry is willing to supply transactions along the positively sloped marginal cost curve. These total marginal costs may be subdivided into costs incurred by merchant banks and those incurred by purchaser banks. Purchasers, on the other hand, through their pooled willingness to purchase transaction services, have effective demands along the line *d*. The intersection of *d* with *c* at point *e* implies an equilibrium price of *p** to facilitate *q** exchanges. In the process of producing an industry output of *q**, merchant banks incur marginal costs in the amount *q***a* and purchaser banks incur marginal costs in the amount *q***d*; and the sum of those two sets of costs is *q***e*. In consideration for transactional services to facilitate *q** exchanges, purchasers are willing to make expenditures in the amount of *q***b* and merchants are willing to make expenditures in the amount *q***c*; the sum of those two revenues streams is *q***e*.

What is of critical importance is that the marginal cost *q***d* of the activities performed by purchaser banks bears no necessary relation to the amount of revenue *q***b* forthcoming from the purchasers with whom those banks have contractual relationships. Similarly, the costs *q***a* associated with the activities performed by merchant banks have no necessary relation to the amount of revenue *q***c* forthcoming from the merchants with whom they have contractual relationships. Nonetheless, the sum of the two revenue streams equals the sum of the two marginal cost streams, *q***e*, and it follows that there must be some particular side payment between a merchant bank and purchaser bank with respect to any particular exchange that will bring the receipts of each bank into equality with the marginal cost it has incurred in providing transactional services to facilitate the exchange.

In Figure 4, *M* bank receives *q***c* of revenue from merchants and must pay over to *P* bank the amount *ac*; and *P* bank receives from its purchasers revenue in the amount *q***b*, which is less than it costs, *q***d*, by the amount *bd*. The side payment from *M* bank, *ac*, precisely equals the deficiency, *bd*.[8]

It is true, of course, that a side payment of *ac* per facilitated exchange from *M* bank to *P* bank is not the only conceivable institutional adjustment, but it

8 By construction, $q^*e = q^*a + q^*d = q^*b + q^*c$; hence, rearranging, $q^*d - q^*b = q^*c - q^*a$. But $q^*d - q^*b = bd$, the revenue deficiency of *P* bank; and $q^*c - q^*a = ac$, the revenue excess of *M* bank. It should be clear that nothing turns on the fact that I have drawn the diagram in such a way that c_P lies above c_M in the range q^* or that d_M lies above d_P in that range. No matter what combination of these relationships exists, as long as the sum of the revenues equals the sum or the costs, then notwithstanding that *P* bank's revenues from its purchasers do not equal its costs, there is some transfer payment between the two banks that will bring revenues into equality with costs for each.

appears to be by far the simplest and the least expensive.[9] Since any redistribution mechanism will itself involve a transaction cost which will serve to raise *C*, the mechanism that minimizes transaction costs is in the interest of all the parties. Since remittance of funds in some amount from *P* bank to *M* bank is an inescapable feature of any payment mechanism of the type under consideration, adjustment of the magnitude of that remittance to achieve the equilibration of costs and revenue clearly appears to be the preferred mechanism.

In summary, one would expect to observe the following behavior in the operation of cashless payment systems: after the purchase transaction between *P* and *M* (1) *M* bank buys the paper from *M* at face value, minus a discount in the dollar magnitude *q*c*, thus bringing revenues of *q*c* into the banking system; (2) *P* bank buys the paper at face value from *M* bank, minus a discount (*q*c* – *q*a*), leaving *M* bank with net revenues *q*a*; (3) *P* bank bills its customer *P* in an amount equal to the face of the paper plus the premium *q*b*, thus bringing revenues in the amount *q*b* into the banking system. Thus in total *P* bank has received revenues in the amount *q*b* + *q*c* – *q*a*. But the first two terms in that expression are equal to *q*e*; and *q*e* minus the third term, *q*a*, is equal to *q*d*, *P* bank's costs.

One important assumption underlies the preceding paragraph: banks participating in the payment system are behaving competitively and charging prices to *P* and *M* corresponding to the bank's marginal costs and. in equilibrium, to their average total costs including the opportunity costs of invested capital. There are two quite distinct reasons why this assumption may not hold in any particular real world context. First, through collusion the banks might have acquired enough market power to be able to charge both purchasers and merchants prices that exceed the banks' cost.[10] I explore the implications of collective action among banks more fully, later in this paper.[11] For the present, I note only that the problem of cartel profit maximization will be complicated by the fact that, in order to maintain an equilibrium number of transactions,

9 The phenomenon discussed in the text occurs in any four-party transaction in which each of two transacting principals is represented by an independent agent or broker, each of whom also incurs costs. The costs of the two brokers must be paid out of the theoretically possible gains from trade between the two principals. Tradition and transaction-cost considerations may require that the selling principal compensate the selling broker and the buying principal compensate the buying broker; yet there may be no equivalence between the height of each principal's demand curve for brokerage services and the costs incurred by his broker. Often a side payment between principals in the form of an adjustment to the underlying sale price will be used to achieve equilibrium. In such a situation the form of the side payment obscures its very existence and also obscures the complexity of the equilibrium that is being attained. Many brokered real estate transactions answer this description. In four-party payment mechanisms, too, a side payment between *P* and *M*, coupled with payment by each *P* and *M* to *P* bank and *M* bank, respectively, in amounts equal to respective bank costs but not to respective marginal utilities of *P* and *M*, is theoretically sufficient to attain equilibrium. That in practice side payments between banks occur instead is strong evidence that higher transaction costs characterize side payments that take the form of price adjustments between the principals.

10 See generally William M. Landes & Richard A. Posner, Market Power in Antitrust Cases, 94 Harv. L. Rev. 937 (1981).

11 See Sec. III infra.

the cartel must increase prices each to merchants and to purchasers in amounts dictated by the slope of their demand curves—amounts that, in all probability, are equal neither in absolute magnitude nor in percentage markup over the competitive price. Hence cartelization of the industry would be comparatively difficult.[12]

The second reason that some degree of market failure might be observed involves the relations between the two sets of banks. Each *M* bank collects transaction paper that must be forwarded for collection to many *P* banks, including some with which that *M* bank will never before have dealt. At that time, *M* bank faces a monopsonistic buyer for each piece of paper. One can imagine a variety of institutional solutions for this problem. Conceivably, *P*'s participation in the payments system could be conditioned on his assuming an obligation to redeem his paper from any bank that presented it to him. Under that arrangement, *M* bank would face a competitive set of bidders for *P*'s paper, but such an arrangement would so increase *P*'s transaction costs that the competitive viability of the payment system, in competition with others, would be in serious doubt. Moreover, if the payment system in question involves a deposit relationship between *P* and *P* bank, accompanied by an understanding that the paper will be debited against *P*'s deposit, *P* bank would nevertheless remain in a significant monopsonistic position: it would have lower float costs and lower default costs because of the security afforded by the existence of the deposit.

12 Assume that credit cards are issued to card holders only by a single bank, *P* bank, which is effectively sheltered from competition by law; and assume that merchants are serviced by a competitive set of merchant banks. Then *P* bank can maximize profits by restricting output to a level *q"* below *q**, at which the total marginal cost curve, *c* in Figure 4, signals the marginal revenue curve (not shown in Figure 4) pertaining to the aggregated curve *d*. But since there must be some particular rate *q"* at which transac tions are contracted, the output restriction implies a higher price in equilibrium to card holders as well as to merchant banks and merchants. An increase in the interchange fee without an increase in card holder fees would result in a decrease in the number of card transactions that merchants were willing to enter without reducing the number that card holders were desirous of entering. This would reduce the aggregate utility of the card system to card holders simultaneously with increasing the utility to card holders of the marginal transaction each was able to enter. Thus *P* bank would be forgoing the opportunity to exploit, through cardholder fees, that higher marginal utility. This pattern would create incentives for card holders to make side payments to merchants to induce additional transactions. Because those side payments must be presumed to involve higher transaction costs. *P* bank would he squandering its monopolistic potential. Assume, more realistically, that credit cards are issued by a group of banks that own the card system as a cooperative venture and share in the profits of the system proportionately to the dollar volume of charge transactions executed by each member's card holders. Now any attempt to exploit merchant banks (and merchants) by increasing the interchange fee is doomed to failure, quite apart from competition from rival payment mechanisms, unless the member banks also act collectively to exploit card holders. If member banks compete actively for card holders, as they would have strong incentives to do, to increase their share of interchange monopoly profits, they will simultaneously dissipate the monopoly profits and create incentives; even stronger than those previously described, for card holders to make side payments to merchants. Equilibrium is attained at zero monopoly profits, needlessly high transaction costs, and a smaller industry than under competition. Cartelization with respect to the merchant's demand function without simultaneous cartelization with respect to the card holder's demand function would not appear to be feasible; and cartelization with respect to both demand functions is made difficult by unusually high information requirements about the relative positions of the two demand functions, in addition to the usual difficulties of policing cheating by cartel members through rivalry for card holders.

In short, if *P* is to be afforded the transaction costs savings associated with having his paper returned to him through one particular *P* bank, and if deposit-based transaction systems, as opposed to pure credit systems, are to be among the set of systems available, *M* bank must have, at the time it acquires paper from its set of merchants, a preexisting understanding governing interbank discount with each bank in the set of participating *P* banks. If the number of *P* banks participating in this system is large, as it often will be, a complete set of bilaterally negotiated agreements would be excessively cumbersome and costly. Some uniform understanding between the set of *M* banks on the one hand and the set of *P* banks on the other would appear to be essential to any cost-effective payment system.

As we shall see, the practical and legal difficulties of bringing into existence such a uniform understanding constitute a significant part of the history of the various payment systems.

II. THE HISTORY OF FOUR-PARTY TRANSACTION VEHICLES

Over the last 150 years, three distinct categories of four-party cashless payment systems have evolved. The check and the bank credit card are heavily used today to facilitate exchanges, and the debit card is increasingly being promoted. This section presents a brief history of the commercial environment in which each of these instruments was introduced and of the practices that developed in conjunction with each of them. By use of the economic theory developed in Section 1, it is possible to uncover previously unrecognized forces in the evolution of these payment systems.

A. The Practice of Paying Checks "At Par"

In the early 1800s the two principal means of payment in commercial transactions were (i) bank notes issued by state banks and (ii) drafts. These two media can be thought of as corresponding to (i) currency and (ii) checks today. Although checks had an early origin,[13] they did not become common until

13 The use of checks in America had its origins in the operation of "the fund at Boston" in 1681. A person could direct the manager of the fund, in writing, to transfer part of his deposit to the credit of another. However, the use of deposit currency, or checks proper, did not become common until a century later. W. E. Spahr stated, in his excellent history of checks, that deposit currency did not develop until after the Revolutionary War, for the following reasons: (1) The colonists had very little specie to deposit. (2) The country was sparsely and deposit banking implies that the inhabitants be in close touch with their banks in order to test the validity of their checks. (3) There was not the requisite security of personal and property rights and confidence in government and banking institutions. Walter E. Spahr, The Clearing and Collection of Checks 38-43 (1926).

after the Revolutionary War.[14] In the years between the demise of the Second Bank of the United States and the Civil War, checks were commonly used as a means of paying local bills only in the nation's commercial centers.[15] City banks encouraged the use of deposit currency because inferior country bank notes of uncertain value tended to drive the sounder city bank notes out of circulation.[16] For the most part, the attempts of the city banks to prevent the discounting of these notes were unsuccessful.[17] During this time, transportation outside the nation's commercial centers was slow, expensive, and often dangerous. Only infrequently did either goods or people travel very far. Markets were predominantly local, and goods consumed in any geographic area usually had been produced there.

In those commercial circumstances, *P* and *M* were almost always residents of the same area. Accordingly, payment media rarely had to be sent beyond the local area. Bank notes, issued by the local bank or banks, circulated through the area and were used in a far greater fraction of transactions than currency isused today.[18] In the larger local transaction, and also in the relatively infrequent long-distance transaction, the draft was the typical medium used.[19]

14 The use of checks for local payments accelerated after the Revolution. There is substantial evidence of the use of checks in the nation's commercial centers before the creation of the first United States Bank in l79l. *Id*. at 43. Spahr estimated the amount of check use in America by examining the relation between deposits and currency in circulation. Deposits passed bank note currency in 1855. *Id*. at 60. In 1867 the public held $l.20 in deposits for every dollar of currency and, by 1872, held $2.00 for every dollar of currency. After 1880 the ratio began a long-term climb; it was twelve to one in 1929. Milton Friedman & Anna Schwartz, A Monetary History of the United States 16 (1963).

15 Federal Reserve Bank of Richmond, Letter No. 4, Mar. 1922, reprinted in Readings in Money, Credit and Banking Principles 377, 379 (Ivan Wright ed. 1926).

16 Broy Hammond, Banks and Politics in America: From the Revolution to the Civil War 549 (1957).

17 However, the banks in Boston, under the leadership of the Suffolk Bank, were able to institute a system that discouraged the discounting of New England Bank notes. *Id*. at 549-56; V. Longstreet, Currency Systems of the United States in Banking Studies 65, 69 (Federal Reserve ed. 1941). See note 45 *infra* and accompanying text.

18 See note 14 *supra*. Bank notes were far more important to country banks, especially those in the southern and western states, than for the city banks. In 1841, "Gallatin pointed out that deposits constituted the principal currency in the larger cities but that country banks could not exist unless they had the right to issue bank notes." Spahr, *supra* note 13, at 63.

19 Although there is a consensus that the draft was the principal means by which a buyer in the country paid a long-distance debt during the early part of the nineteenth century, there is disagreement about the duration of the practice. Thatcher C. Jones, Clearing and Collections 172-74 (l93l); Testimony on Par Collection of Checks: Hearings on H.R. 12379 Before the House Comm. on Banking and Currency, 66th Cong., 2d Sess. (1920), indicates the importance of the use of drafts up until the l890s. But Claudius B. Patten, writing on the mid-1880s, stated that although the use of drafts was common thirty to forty years previously, "Nowadays no country trader, no matter whether he is located in Deadwood or St. Augustine, thinks he is in fashion unless he 'pays' his New York or Boston bills by sending there his individual checks on his local bank, which gets all the advantage of his deposit until the checks come around for collection from the city banks, which have given their dealers immediate credit for them, and made no charges for their collection." Claudius B. Patten, The Methods and Machinery of Practical Banking 1100-01 (11th ed. 1902).

If *P* became indebted to *M*, who resided in a distant place, *P* would execute payment by purchasing a draft made payable to *M* as payee. His local *P* bank would prepare a draft instructing *M* bank in *M*'s geographic vicinity to make payment to *M* in the amount of the indebtedness. For this service, *P* would pay a very substantial fee in comparison with present day transaction costs. In the terminology of the day, *P* was said to "purchase exchange" from *P* bank.[20] The draft thus obtained would then be sent through the mail, usually by *P* bank but perhaps by *P* himself, addressed either to *M* bank or to *M* himself. If sent to *M*, the draft would be presented by him to *M* bank for payment; or if sent to *M* bank, the draft would be held while notice was transmitted to *M* that funds were available to him at *M* bank.

This transaction satisfied the obligation of *P* to *M* but created a new indebtedness on the part of *P* bank to *M* bank. This interbank indebtedness might then be settled in any of several ways. Settlement was simplest if *P* bank customarily maintained a positive balance with the remote *M* bank; and the existence of such a correspondent relationship between *P* bank and *M* bank would have been a sufficient reason to select *M* bank as drawee of the draft in *M*'s favor. If no such balance was maintained, *P* bank might now settle its indebtedness by issuing and mailing yet another draft, payable to *M* bank, to some third bank with which it did maintain a balance, that third bank being selected because it was geographically close to *M* bank. Alternatively, if *P* bank maintained no such balance in *M* bank's vicinity, *P* bank would now be obligated physically to transport to *M* bank a mutually acceptable form of currency. In either event, the cost of the transaction was substantial: the costs of shipping bank notes or gold were high, as were the opportunity costs of maintaining non-interest-bearing balances at distant locations. It was to cover these costs that *P* paid to *P* bank a substantial service charge in addition to the face amount of the draft.[21]

In 1864 Congress passed the National Bank Act,[22] reinstituting the rivalry between state and national banking systems that had existed during the nation's first half century. Federal taxes were levied on bank notes issued by

20 The fee charged by *P* bank was referred to as the "charge for exchange" or, often, "exchange." The amount of this exchange varied greatly with the circumstances of the case, but generally speaking it was large enough to cover the cost to *P* bank of sending currency to *M* bank, including the transportation charges, insurance, and interest on the money in transit. Federal Reserve Bank of Richmond, *supra* note 15, at 380.

21 The average price of southern and western exchange on New York markets in 1859 was estimated to vary from 1 to 1.5 percent. After 1890 the charges varied from one-tenth to one-fourth of 1 percent. Spahr, *supra* note 13, at 102.

22 In 1863 Congress passed "An Act to provide a national Currency, secured by a Pledge of United States Stocks, and to provide for the circulation and Redemption thereof." Act of Feb. 25, 1863, ch. 58, 12 Stat. 665. The 1863 law was replaced by the Act of June 3, 1864. ch. 106, 13 Stat. 99. This Act established the National Banking System and is commonly known as the National Bank Act.

state banks in an endeavor to drive the notes, and perhaps the banks, out of existence.[23] Although the 1864 Act required that national banks maintain reserve deposits, it permitted a large fraction of those reserves to be held as deposits in designated "reserve banks" in various major cities; and, because drafts could be issued against these reserves, the national banking system became instrumental in the payments system.[24]

The era was one of rapid technological change in both transportation and communications. The railroads, waterways, and post roads expanded rapidly, frequently under the spur of government subsidies, and the telegraph was invented and deployed. These changes tend to explain the increase in use of transactional paper relative to currency, but it is less clear why the use of checks relative to drafts also increased very rapidly during this period.[25] When a check was used to pay a distant payee, *P*, having a positive balance with *P* bank, sent the instrument (usually by mail) to *M*, who presented it to *M* bank for collection. Then *M* bank accepted the instrument for collection and might or might not credit *M*'s account with *M* bank for the amount of the check before collection had been achieved.[26] The instrument was started by

23 A tax of "ten per centum on the amount of notes of any state bank, or state banking association" was levied by Congress. Act of Mar. 3, 1865, ch. 78. § 6, 13 Stat. 484. One year later the tax was reenacted by Congress with a more extended application. Act of July 13, 1866, ch. 184, § 9, 14 Stat. 146. The Supreme Court upheld the constitutionality of the tax in Veazie Bank v. Fenno, 75 U.S. (8 Wall.) 533 (1869). Because of widespread evasion of the law by banks, corporations, and municipalities, Congress repealed the Act and substituted a more comprehensive prohibition. Act of Feb. 8, 1875, ch. 36, §§ 19-21, 18 Stat. 311. The tax, which was intended not only to eliminate state bank notes but also to force the state banks to become national banks, did not achieve the second purpose. State banks managed to survive by increased reliance on deposit currency. See Hammond, *supra* note 16, at 753. Although the tax initially caused many banks to become national banks, the decline (as measured by the decreasing size of state and private bank deposits) ceased in 1867. By 1871 the deposits in nonnational banks had expanded to the point where they equaled the deposits of the national banks. See Friedman & Schwartz, *supra* note 14, at 19. See also Kenneth W. Dam, The Legal Tender Cases, 1981 Sup. Ct. Rev. 367, for a treatment of the causes and consequences of the legislation in this period.

24 Country banks used their reserves as a means of clearing their checks without paying remittance charges. After the banks in New York City started charging for the collection of these out-of-town checks, the reserve balances were transferred to other cities. Spahr, *supra* note 13, at 110-11; Charles F. Dunbar, The Theory and History of Banking 50 (4th rev. ed. 1922).

25 "By taxing State bank notes out of existence in 1865, a vacuum was created which gave an added impetus to the use of deposit currency. Other factors which were responsible for the increasing use of deposit currency, and consequently checks, were the inelastic note currency, better means of communication, the cheap and uniform postage rates, and the denser population." Spahr, *supra* note 13, at 84. Spahr explains the greater use of out-of- town checks in the following manner, "As the banks grew in numbers and the use of checks in payment of foreign (out of town) bills became more general, the banker found he could charge the collecting bank a maximum rate with less compunction than he could charge his depositor a minimum rate on drafts, and so he encouraged the use of the check." *Id.* at 103. These comments leave unexplained why *P* was expected to pay for exchange but *M* bank was expected to pay when checks were used.

26 Competition soon forced banks into the practice of crediting immediately the uncollected checks to the depositor's account and paying interest on those uncollected funds. Spahr, *supra* note 13, at 110.

M bank on what was often a circuitous journey from one bank to another until through some series of correspondent relationships it arrived at *P* bank.[27] The check was accepted by *P* bank and debited against *P*'s account. At this point *P* bank again faced the problem of making payment to *M* bank, just as when drafts were used. Again, its costly alternatives were the actual transport of currency or the maintenance of geographically dispersed balances against which a draft in favor of *M* bank could now be issued.

To obtain revenues, *P* bank might have levied a service charge against *P*'s account and made remittance to *M* bank in the full face amount of the check; but this was not the custom. Rather, it was customary to make remittance to *M* bank in an amount less than the face of the check, the discount being called an "exchange charge," a term that reflected the functional similarity of the charge to the prepaid service charge characteristically imposed on *P* in the earlier period when a draft was issued on his behalf. The preservation of that term, however, tended to obscure the important fact that the direct economic incidence of the service charge had been shifted—initially to *M* bank, or to some intermediate bank in the chain which might be willing to absorb the charge. but ultimately to *M*.

Early descriptions of the checking system suggest that the contemporaneous view in the banking community of this shift in incidence was that it reflected an understandable conflict of interests between *P* bank and *P* on the one hand and *M* bank and *M* on the other.[28] But that explanation fails for two reasons. First, the conflict of interests had been present no less during the earlier period when drafts were the predominant transaction vehicle; and old causes cannot explain new effects. Second, the explanation attributes a widespread and persistent pattern of behavior to an erroneous perception, for it implicitly assumes that the checking system could attain equilibrium without regard to the proportion in which banking costs were imposed on *P* and *M* so long as all costs were borne by them in combination. To the contrary, as I argued in Section 1, equilibrium in the level of checking services demanded and supplied is possible only with some specific distribution of costs between *P* and *M*.

If the shift in incidence reflected rational business behavior, as I prefer to think it did, then it had to reflect either a change in the relative demands of purchasers and merchants for checking services or changes in the relative costs of *P* bank and *M* bank in providing them. Several contemporaneous developments support the inference that such shifts actually occurred.

---

27 One check traveled 1,500 miles and passed through eleven banks in an attempt to avoid remittance charges. James C. Cannon, Clearing House Methods and Practice 74-78 (1900), reprinted in U.S. National Monetary Commission, Clearing Houses and Credit Instruments 70-74 (Publications of the Nat'l Monetary Comm'n No. 6, 1910). See also Spahr, *supra* note 13, at 105.

28 Spahr, *supra* note 13, at 18. Current explanations also use conflict-of-interest explanations, for example, Hal Scott, The Risk Fixers, 91 HARV. L. REV. 737 (1978).

The advent of faster and cheaper transportation and communication had two consequences for the supply costs of transactional paper. First, it reduced the banking system's aggregate direct costs of processing checks and, when necessary, transporting currency. Second, because they tended to convert local markets into regional and national markets, these cost reductions greatly increased commercial transactions between remote parties. This increase in the volume of distant transactions enabled banks to exploit scale economies in maintaining balances at distant locations; for, given the law of large numbers, higher turnover velocities in those balances could be achieved with disproportionately small increases in the magnitude of the balances. This factor, too, must have contributed to a reduction in average cost per transaction.

In addition, although under the draft system *P* contributed substantially to bank revenue by purchasing "exchange," those transactions imposed large indirect costs on *M*: the cost of the float during the slow process of paper interchange and the cost associated with the risk of default. In addition to the reductions in direct cost brought about by better transportation and communication, these indirect costs to *M* would also be significantly reduced by shortening the period of float, by providing cheaper access to credit references, and by reducing the costs of collecting delinquent obligations. Hence, even if there had been no reduction in aggregate direct costs, the redistribution of those direct costs toward *M* might well have been necessary to attain equilibrium in view of the reduction of *M*'s indirect costs.

Finally, the widespread emergence of clearinghouses also significantly reduced direct costs and accelerated the process of interchange, further reducing float costs.[29]

For some or all of these reasons it seems to have been necessary for the industry to redistribute the direct costs of the checking system away from *P* and toward *M* so that the market for transactional paper could equilibrate. That need may itself best explain the relatively sudden displacement of the draft by the check. A new and less familiar instrument, the check was accompanied by fewer customs and fixed expectations than the more familiar draft. And the check, although very similar to the draft in most respects, passed through the hands of the four parties in a different sequence, a sequence that tended to enhance monopsonistic position of *P* bank as a buyer of paper.

As Figure 4 demonstrates, if the level of total banking costs (and therefore the values of p^* and q^*) changed significantly, then no change in the aggregate demand curve of *P* and *M* would be necessary to change the relative magnitudes of their individual demand levels for use of a payment system. It is well

29 See generally Cannon, *supra* note 27. The first clearinghouse was established in New York City in 1853. During the following five years clearinghouses were established in Boston. Philadelphia, Baltimore, and Cleveland. By the mid-1870s clearinghouses were established in most of the leading cities in the United States. In 1899, there were 31 clearinghouses in the United States. Dale H. Hoffman & Melvin Miller, Origin and Development of Charges for Banking Services 10-14 (1942).

established that from the Civil War to the end of the nineteenth century p^* fell by a considerable amount and q^* increased enormously.[30]

The clearinghouse seems to have had consequences beyond mere reduction of costs to the banking system. With increasing urbanization of the nation, many banks found themselves in cities served by many other banks. The local clearinghouse—at which each bank in its role as *M* bank would transfer to every other bank in its role as *P* bank a bundle of checks, packaged and tallied in advance—had enormous potential for reducing the costs of the payment system by expediting both presentment and remittance. Interbank debits among clearinghouse members could be netted out on the books of the clearinghouse; and actual payment, usually made to the clearinghouse, was necessary only intermittently to the extent that an individual bank's presentment over a period of time had aggregated more or less than the aggregate, over the same period, of its remittance obligations.

Clearing arrangements were negotiated not only among banks in individual urban areas but also between banks in widely separated urban areas. These intercity arrangements were often bilateral agreements by which one large bank in the first city would accept for forwarding to all other banks there checks gathered in the second city by the other large bank from all other banks located there.

These clearing arrangements were significant because they both reduced the cost per item substantially and encouraged standardization. Because of the large number of items involved and because cost reductions depended heavily on use of routinized procedures for assembling the items in batches and tallying the totals for the items in each batch, it was highly desirable that every item be susceptible to handling in the same routinized way.[31] If different exchange charges were to be charged on different items by different *P* banks—charges not appearing on the instruments—handling procedures would be complicated.

Moreover, many banks were indifferent whether exchange charges were low or high or even made at all. The typical bank presented to other banks about the same volume of items as were presented to it; and for such a bank the aggregate of exchange charges represented a wash. The increased administrative cost of accounting for different exchange charges on different individual items constituted a useless cost for such a bank. Therefore, there was a strong incentive to standardize such charges, and fixing them at zero was an obvious and entirely acceptable form of standardization.

---

30 Compare Wright, *supra* note 15, at 380-81.

31 Albert Gallatin first proposed establishing a clearing system in 1841 as a means of reducing the costs of exchanging checks and notes. See Hammond, *supra* note 16, at 705-07; Spahr, *supra* note 13, at 79-82.

For these reasons, many banks agreed to handle each other's items "at par"—that is, to make no exchange charges. For similar reasons, many organizations required their members to remit at par on all items sent through the clearing arrangement.[32]

An exchange charge equal to zero obviously has no unique potential for cost-reduction; any uniform exchange charge would have facilitated routinized processing. Any advantage of a zero price over others is rooted less in economics than in psychology.[33]

Parties to individual items on which varying amounts of exchange would be charged when they reached *P* bank were at a disadvantage in competing with parties to items eligible for routinized clearance. Clearance mechanisms tended to get a check from *M* bank to *P* bank via quite direct paths, but items on which exchange charges were due tended to follow slow and circuitous routes.[34] Each bank would prefer to transfer the item to another bank with whom it had negotiated a bilateral arrangement to remit at par than to send to *P* bank, which would impose exchange charges. Consequently, both float and handling costs were relatively greater for items with nonstandardized exchange.

Notwithstanding the advantages of uniform (perhaps uniformly zero) exchange charges, a very large number of banks strenuously resisted remitting at par. The banks that continued to charge exchange into the twentieth cen-

---

32 In 1899 the banks of Boston organized a system for the collection of country checks. The Boston Plan was intended to force all banks in New England to clear checks at par. The plan resulted in 97 percent of the checks in New England being collected at par. Under the Boston Plan the cost of collection was reduced from a rate which varied from $1.00 to $1.50 per thousand dollars to a charge of six or seven cents per thousand. Spahr, *supra* note 13, at 128. See Federal Reserve Bank of Richmond, *supra* note 15, at 382-83; note 25 *supra* and accompanying text.

33 See Thomas C. Schelling, The Strategy of Conflict 67-80 (1960).

34 See Spahr, *supra* note 13, at 103-08. See also note 27 *supra* and accompanying text. In the political arena, arguments of doubtful substance were built on the existence of these circuitous routings. Because such routings tended to add to the number of items (and dollar volume of items) outstanding at any point in time, they increased the float—the number of dollars shown as additions to the deposits of *M* bank but not yet deducted from the deposits shown on the books of *P* bank. This phenomenon results in an over statement, in the aggregate, of deposits in the banking system. Since the aggregate of loans that the banking system is able to make is a percentage of deposits, anything that increases the float increases the money supply and tends to have inflationary effects. The increase in the mean money aggregates would represent a one-time event and would be of doubtful significance, but to the extent that the float is less stable than genuine deposits, a large float might also tend to destabilize the money supply. Banks that did not clear at par were criticized for causing these undesirable macroeconomic effects. Slow and circuitous clearance of checks is also undesirable from the standpoint of banking policy because it facilitates the practice of "kiting"—the deliberate manipulation by an individual of deposits and checks outstanding against nonpar banks—and practices were criticized on this basis too. Although this attack may have had more substance than the money supply attack, both confuse the desirability of standardization with that of par clearance. Spahr, *supra* note 13, at 105-08; Federal Reserve Bank of Richmond, *supra* note 15, at 384-89. See note 23 *supra* and accompanying text.

tury were almost without exception, small banks in isolated agricultural communities. For the banks that adhered to this practice, revenue in 1964 from exchange charges constituted about 10 percent of total current operating revenue, and the percentage was higher for the smaller institutions among the group.[35] It seems likely that in the late 1800s and early 1900s, when the nonpar controversy was at its height, this form of income was even more important to the small country bank.[36]

There are at least two possible explanations of how these rural banks benefited from charging exchange. One is that, even though they charged exchange in their role as *P* bank, they managed to collect at par in their role as *M* bank. No doubt this explanation is at least partly correct, for banks that did not remit at par were not, for that reason alone, prohibited from forwarding for collection items drawn on banks that did remit at par via a correspondent bank through the Federal Reserve clearing system, and the same may have been true of some earlier, private clearance systems. But because remittance at par, at least generally, was a reciprocal practice, it seems unlikely that this was the whole explanation. Moreover, although this hypothesis tends to explain why some banks clung to the practice and might, when coupled with another factor I address hereafter, tend to explain why the practice was most common forbanks in isolated communities, it does not explain why the practice should have been confined so largely to isolated agricultural communities, rather than, for example, mining communities.

A different factor must have been at work. The amount of exchange charged was customarily a percentage of the face value of the item. But a minimum charge, often ten cents, was charged on all items having a face amount of $100 or less, and $100 was a large sum then. A bank benefits from charging exchange if, notwithstanding that its aggregate dollar volume of remittances roughly equals its collections, a larger number of small items are presented to it than it presents to other banks. In isolated agricultural communities, the receipts of the farmers, who constituted the rural depositors, probably took the form of several large payments at harvest time. On the other hand, farmers more nearly resemble nonfarmers in their purchase patterns, for they engage in personal consumption and the purchase of farm supplies throughout the year. And, of course, the magnitude of most individual purchasers must be much smaller than the magnitude of the small number of income items. Although apparently no data exist that would constitute hard evidence for this hypothesis, it is the only explanation that enables me to make sense of the available information about the nonpar controversy.

35 Paul F. Jessup, The Theory and Practice of Nonpar Banking 48 (1967).

36 "In many instances throughout the South the exchange revenue of the small or country bank constituted considerably more than half of the bank's income." Federal Reserve Bank of Richmond, *supra* note 15, at 391.

Why nonpar practices tended to be confined to small isolated communities is more obvious. A situation in which one or more nonpar banks occupied the same market with one or more par banks is inherently unstable. It had always been an unambiguous understanding about any bank's obligation on a check that payment had to be made at full face value if the check were presented for payment at its banking premises. If there was a par bank in the same areas as *P* bank, *M* bank would forward items drawn on nonpar *P* bank to that neighboring bank so as to avoid exchange costs; and the neighboring bank would present such items at *P* bank's premises. Hence, the conversion from nonpar to par of any one bank in an area usually led to the conversion of all in the area. Nonpar banking thus survived primarily in isolated communities able to support only one, or a few, banks. However, in the early twentieth century it was Federal Reserve pressure, not competition, that reduced the practice of charging exchange to a trivial level; where the practice survived it was state legislation, not monopoly enclaves, that sheltered it.

After the monetary panic of 1907, a national monetary commission was appointed to study the American banking system.[37] Its report led to the passage of the Federal Reserve Act in 1913.[38] This legislation, its subsequent amendments, and the practices and rules of the Federal Reserve Board, which the legislation created, eventually tipped the balance in favor of par clearance in the United States. It was not obvious from the initial legislation that this outcome would result, nor is there any reason to believe that the practice of nonpar banking particularly concerned either the National Monetary Commission or the Congress of l9l3.[39]

The key provisions of the Federal Reserve Act were sections 13 and 16. Section 13 initially read, in part:

> Any Federal reserve bank may receive from any of its member banks … deposits . . . or, solely for exchange purposes, may receive . . . checks and drafts upon solvent member or other Federal reserve banks, payable on presentation.[40]

37 Act of May 30, 1908, ch. 229, Pub. L. No. 169, §§ 17-20, 35 Stat. 546, 552.

38 Federal Reserve Act, ch. 6, Pub. L. No. 43, §§ 1-30, 38 Stat. 251 (1913).

39 The National Monetary Commission did not make any specific recommendations about exchange charges. Section 16 of the Federal Reserve Act only prohibited member banks from charging other members remittance charges. Member banks were allowed to charge their customers the actual cost of collection.

40 Federal Reserve Act, ch. 6, Pub. L. No. 43, § 13, 38 Stat. 263 (current version at 12 U.S.C. § 342 (1976)).

Section 16 read, in part:

> Nothing herein contained shall be construed as prohibiting a member bank from charging its actual expense incurred in collecting and remitting funds, or for exchange sold to its patrons. The Federal Reserve Board shall, by rule, fix the charges to be collected by the member banks from its patrons whose checks are cleared through the Federal reserve bank and the charge which may be imposed for the service of clearing or collection rendered by the Federal reserve bank…[41]

Section 16 is silent on the practices of nonmembers. It preserves the right of members to impose costs on their check-writing depositors and implies obliquely that language elsewhere in the Act might be read to curtail member *P* bank's ability to charge exchange to *M* bank; but no curtailing language is to be found elsewhere. The power vested in the Reserve Board to standardize fees for clearance or collection at a level other than zero has never been exercised.

More generally, the Act provided that the Federal Reserve Board would establish a check clearance system throughout the United States, each federal reserve bank being required to act as a clearinghouse for member banks in its region. After establishing this system, the Fed began to establish more pervasive clearing mechanisms. Funds for the clearance system were available, for the Act also required member banks to deposit substantial reserves with federal reserve banks in accounts bearing no interest.[42] Deposits, however, were invested in government securities; and the investment yield constituted a very substantial source of funds to the system. It seems clear that the clearance systems established by the Fed were largely subsidized by these earnings. Although member banks did not receive a "free" clearing system—the forgone investment yield on their reserve deposits paid for it—the Fed clearing system was available to members at a price included in the sunk cost of maintaining the expired reserves. The alternatives (to continue using private clearinghouses or to establish a new, private, interregional clearinghouse) would have required that member banks bear the full system costs in addition to the cost of maintaining reserves with the Fed. Accordingly, the economic incentives for member banks to use Fed clearing mechanisms were strong.

41 Federal Reserve Act, ch. 6, Pub. L. No. 43, § 16, 38 Stat. 265, 268 (1913). The only amendment made to the quoted portion of the section is the name of the Federal Reserve Board. The second sentence quoted now reads, "The Board of Governors of the Federal Reserve System…" Act of Aug. 23, 1935, ch. 614, § 302(a), 12 U.S.C. § 360 (1976).

42 Section 9 of the Federal Reserve Act specified the reserve requirements of member banks. The requirements were substantially lowered by the Act of June 21, 1917, ch. 32, Pub. L. No. 25, § 10, 40 Stat. 239. Member banks in central reserve cities were required to maintain reserves of 18 percent against demand deposits (decreased to 13 percent) and 5 percent against time deposits (decreased to 3 percent). Member banks in reserve cities were required to carry reserves of 15 percent against demand deposits (decreased to 13 percent). The reserves of country banks were fixed at 12 percent for demand deposits (decreased to 7 percent) and 5 percent for time deposits (decreased to 3 percent). The reserve requirements were lowered to stimulate membership in the Federal Reserve System. See Federal Reserve Bank of Richmond, Letter No. 5, Apr., 1922, reprinted in Wright, *supra* note 15, at 391-404.

The incentive for member banks to use the Fed's clearance system, coupled with the Fed's requirement that member banks remit at par against items presented to them through the clearance system, served as a significant direct force in the adoption of clearance at par by member banks. This same force operated, albeit indirectly, on nonmember banks. Member banks were allowed to forward through the system for collection not only checks drawn on other member banks throughout the nation but also checks drawn on such nonmember banks as had agreed to remit at par. In order to identify for member banks those nonmember banks whose checks could be sent through the Fed clearance system, the Fed began regularly to publish the "par list," a complete state-by-state list of all nonmember banks that had agreed to remit at par. In addition, from the beginning of the system nonmember banks could use the Fed clearing system by forwarding acceptable items through correspondent banks that were member banks; but in this context, too, a check drawn on a bank not on the par list was not an acceptable item. Such checks had to be cleared outside the system and were denied the benefits of subsidized clearance.

In 1916 Congress amended section 13. Because the Act initially authorized any federal reserve bank to "receive ... for exchange purposes …checks and drafts upon … member or other Federal reserve banks," some doubt existed whether checks on nonmember banks could be received.[43] The clause was amended to read: "Any Federal reserve bank…solely for purposes of exchange or of collection, may receive…checks and drafts, *payable upon presentation within its district*. . . ."[44] Congress thereby made clear that the federal reserve banks were authorized to accept from their member banks checks drawn on nonmember banks.[45]

Notwithstanding these various enticements, many banks refused to remit at par and stayed outside the federal clearance system.[46] To entice or coerce more banks into its clearance system, the Fed in 1916 made its system mandatory for all member banks with respect to items drawn on them, but the system remained voluntary with respect to items forwarded by them.[47] And nonmember banks on the par list were permitted to ship funds for the purpose of clearance to the Fed at the Fed's expense. Thus a subsidy was employed to expand the par list of nonmembers.

---

43 Federal Reserve Act, ch. 6, Pub. L. No. 43, § 13, 38 Stat. 263 (current version at 12 U.S.C. § 342 (1976)).

44 Act of Sept. 7, 1916, ch. 461, Pub. L. No. 270, 39 Stat. 752 (current version at 12 U.S.C. § 342 (1976)) (emphasis added).

45 Federal Reserve Bank of Richmond, *supra* note 42, at 402.

46 In 1916 the number of member banks actually underwent a slight decline from 7,631 to 7,614. Spahr, *supra* note 13, at 218.

47 Federal Reserve Bank of Richmond, *supra* note 42, at 400.

In 1917 Congress further amended section 13 by adopting the "Hardwick Amendment," which added the language, "Nothing. . . in this Act shall be construed as prohibiting a member or nonmember bank from making reasonable charges, to be determined. .. by the. .. Board, but in no case to exceed 10 cents per $100 or a fraction thereof, based upon the total of checks and drafts presented at any one time, for collection or payment . . but no such charges shall be made against the Federal reserve banks."[48] In its annual report for 1917, the Fed said of the Hardwick Amendment and its legislative history:

> An effort was made, in the interest of some member and non-member banks to amend the Act by providing for a standardized exchange charge, not to exceed one-tenth of 1 percent, to be made by member banks against Federal reserve banks for checks sent for collection. It was not successful, and the Act as finally amended provides that a member or non-member bank may make reasonable charges to be determined . . . by the . . . Board . . .; but no such charges shall be made against the Federal reserve banks." The Attorney General has been requested to give his opinion as to whether this proviso applies to non-member banks. An affirmative opinion will make possible the establishment of a universal par clearing system, but if, on the contrary, it should be held that the proviso applied to member banks only, the further development of the collection system will necessarily be slow, and in the absence of further legislation will depend upon the voluntary action of many small banks.[49]

This comment is noteworthy in two respects. First, it tends to support the view that standardization of exchange charges was seen as a means, alternative to par payment, to facilitate the clearance process. Second, it reveals that the Fed as early as 1917 perceived that the last twelve words of the amendment, if "favorably" interpreted by the attorney general, would be used to coerce a general abandonment of any exchange charges—making "possible the establishment of a universal par clearing system"—and thus achieving standardization of a special kind.[50]

In 1918 the Fed dropped all per item service charges for using its clearance system. It also began operating a leased telegraph system (the "Fed Wire") between all federal reserve banks, the Fed, and the Treasury. The use of the Fed Wire was made available to member and par-list banks to adjust clearing balances. Despite this additional carrot, there remained at the end of 1918 about 20,000 nonmember banks, half of which also remained off the par list.[51]

---

48 Act of June 21, 1917, Pub. L. No. 25, 40 Stat. 234 (current version at 12 U.S.C. § 342 (1976)).

49 Excerpt in Federal Reserve Bank of Richmond, *supra* note 42, at 406.

50 *Id.*

51 At the end of 1918 there were 8,692 member banks of the Federal Reserve System and 10,305 nonmember banks remitting at par, and 10,247 nonmember banks not on the par list. Federal Reserve Bank of Richmond, *supra* note 42, at 407.

In 1918 the Fed succeeded also in obtaining from the attorney general an opinion that in effect prohibited precisely what the Hardwick Amendment seems, at first glance, to have permitted. Focusing on the last few words in the Amendment, the attorney general ruled that the federal reserve banks were prohibited by law from paying, even in the sense of passing on, exchange charges in the course of the clearance process.[52] Since, in the period under discussion, the system would not accept items drawn on nonmember banks not on the par list, the clause, even thus interpreted, would appear to have been inconsequential. But the Fed made it of consequence in 1919, adding substantially to the number of banks on the par list by introducing a new coercive device.

It began to accept for clearance items drawn on nonpar banks and then to demand that they be paid at par. If that request was refused, as it often was, the local reserve bank gathered up the checks of the nonpar bank and presented them at the bank's premises ("at the window"), demanding payment in full in currency.[53] This tactic proved to be very powerful while it was available to the Fed. It has always been regarded as the legal obligation of *P* bank to *P* to pay in full on demand if an item was presented at the window;[54] only with respect to items presented through the mails had banks asserted the right to remit at discount. The batch presentation of checks in the manner described often required more currency than the bank had in its vault; yet if payment in full was not made, the checks could be returned to the depositor dishonored, placing the drawee bank in violation of its contractual obligation to its customer. Through this tactic the Fed succeeded in forcing many recalcitrant banks onto the par list.[55]

Commenting on its endeavors in its annual report for 1919, the Fed said:

> [The] proviso in Section 13... has been constructed by the Attorney General…as meaning that a Federal reserve bank cannot legally pay any fee to a member or non-member bank for the collection and remittance of a check. It follows, therefore, that if the Federal reserve banks are to give

52 *Id*. at 408; Spahr, *supra* note 13, at 234-35.

53 Federal Reserve Bank of Richmond, Letter No. 6, May 1922, reprinted in Wright, *supra* note 15, at 410-12. This tactic of going to the window of the noncomplying bank aid demanding fill payment had been used before as a means of achieving a system of par clearance. The Suffolk Bank System in the 1820s (see Justice Story's decision in Suffolk Bank v. Lincoln Bank, 22 Mass. 106 (1827)) and the Country Checks Department or the Boston Clearing House in the l890s (see note 32 *supra*) both used the same tactic to force par clearance. The Suffolk Bank System was primarily designed to prevent the discounting of bank notes. See Spahr, *supra* note 13, at 73-78, l26-29; Federal Reserve Bank of Richmond, *supra* note 15, at 379.

54 See Spahr, *supra* note 13, at 103-04.

55 In 1919 the number of par banks increased from 18,905 to 25,486 and the number of nonpar banks decreased from 10,191 to 4,015. Federal Reserve Bank of Richmond, Letter No. 5, *supra* note 42, at 410.

> the service required of them, under the provisions of Section 13 they must, in cases where banks refuse to remit for their checks at par, use some other means of collection, no matter how expensive.
>
> The action of the various Federal reserve banks in extending their par lists has met with the cordial approval of the Federal Reserve Board, which holds the view that under the terms of existing law the Federal reserve banks must use every effort to collect all bank checks received from member banks at par. Several of the Federal reserve banks are now able to collect on all points on their respective districts at par, and new additions to the other par lists are being made every day. The board sees no objection to one bank charging another bank or a firm or individual the full amount provided in Section 13 of the Federal Reserve (10 cents per $100) and has not undertaken to modify these charges, but the Act expressly provides that no such charge shall be made against the Federal reserve banks.[56]

The legality of this practice by the Fed was challenged in the courts. While the cases were making their way to the Supreme Court, a number of states, mostly in the rural Southeast, passed legislation providing that a state bank should not be deemed to have dishonored a check—that is, to have violated its obligation to its depositor—if it refused to accept the check merely because exchange would not be paid.[57] The constitutionality of these state statutes was also challenged on preemption grounds.[58]

The two groups of cases made their way to the Supreme Court, which in 1923 held, first, that in the absence of the state statue prohibiting its practice, the Fed was authorized to employ the tactic of making presentment at the drawee bank window[59] and, second, that the state statutes prohibiting the practice were also constitutional.[60] Thus nonpar banking continued to be sheltered in those few states that chose to adopt such statutes but substantially disappeared elsewhere. At the end of 1964, there were 1,547 nonpar banks in fourteen states, but their deposits accounted for only about 2 percent of total deposits in FDIC-insured institutions.[61] On April 1, 1980, there were only fifteen nonpar banks left in the United States.[62] All these banks were located in Louisiana. By September 1980 all but one of these had become par banks.[63]

---

56 Federal Reserve Bank of Richmond, *supra* note 53, at 4125-16.

57 For an excellent discussion of the specific statutes see Spahr, *supra* note 13, at 251-54.

58 *Id*. at 256-90.

59 American Bank & Trust Co. v. Federal Reserve Bank of Atlanta, 262 U.S. 643 (1923).

60 Farmers & Merchants Bank v. Federal Reserve Bank of Richmond, 262 U.S. 649 (l923).

61 Jessup, *supra* note 35, at 23.

62 Federal Reserve System, Memorandum on Exchange Charges (September 1, 1980).

63 *Id*.

Thus the role of the interchange fee in the process of check clearance, a commercial context in which an unregulated market solution might have been expected to work reasonably well and to yield instructive results, was aborted and continues to be suppressed by a mixture of subsidies and coercion by the Federal Reserve System.

B. Bank Credit Cards and the Interchange Fee

About a century passed between the date the check gained common acceptance and the date another four-party payment instrument—the bank credit card—was introduced. The precursors of the bank credit card were the retail merchant's open book account and later the travel and entertainment card.

For centuries merchants have extended short-term, interest-free credit to customers whose patronage is highly valued. The shopping behavior of customers varies widely, and those behavioral differences make transactions with some customers more profitable for the merchant than transactions with others. A customer whose own time costs are high will tend to shop regularly at a particular retail outlet because of its geographic proximity to his other activities, and he will tend to shop when it is convenient for him rather than waiting for occasions when merchandise is on sale. He will tend to shop on fewer occasions and buy a larger number of items on each occasion. He will consume less time of sales personnel because he is attempting to save his own time, and he will be able to decide more quickly because he conceives his quest to be locating the items he wants rather than making closely balanced trade-offs with reference to price. Finally, he will tend to buy higher-priced items, which are likely to carry higher percentage markups and are certain to carry higher absolute dollar markups.

There is a strong although not perfect correlation between customers with high time costs, high incomes, and high wealth positions, so the default risk of extending credit to such customers is also relatively low. For all these reasons merchants have long used the selective extension of open book credit as a competitive tool by which to attract and retain the patronage of such customers.

The customer to whom open book credit was extended, having purchased on various occasions during the month, received by mail at the end of the month a bill in the face amount of his purchases; soon thereafter, he would remit payment by mail. On the average mid-month purchase, the merchant was absorbing the cost of capital for about three weeks. The merchant thus remitted to these customers in a fairly direct way part of his cost savings attributable to their shopping behavior; he also conferred minor indirect cost savings by reducing the customer's need to carry cash on his person.

Open book credit well served the parties affected while travel outside one's home community was relatively infrequent. After World War II, the frequent traveler was likely to have a high income and high time costs and therefore to have been extended open book credit in his own community; but away from home he could not readily be identified at the point of sale. He could carry large amounts of cash, but the risk of loss was substantial. Traveler's checks were an alternative, but they involved high time costs because they required the traveler, first, to visit the bank before departing and, second, to predict with reasonable accuracy how much money would be needed during the trip or to make another journey to the bank on return to redeem the excess checks, or to leave funds tied up on a non-interest-bearing certificate until a later time when the traveler's checks might be used. A second alternative—attempting to cash personal checks at one's destination—involved tediously presenting identification at a moment when time costs were likely to be greatest: not infrequently, the attempt was humiliatingly unsuccessful. From the standpoint of the merchant located at the traveler's destination, the situation was also unsatisfactory—if the merchant could easily identify the traveler as a creditworthy consumer with high time costs, he would be only too happy to extend to the traveler the same credit facilities extended to comparable local customers.

The first commercial response, in the early 1950s, to this obvious transactional need was the travel and entertainment (T&E) card, notably the American Express card and the Diner's Club card. The issuing organization signed up merchants across the country of the type frequently patronized by travelers: hotels, resorts, restaurants, and a relatively small number of prestigious merchandise outlets. After investigating an applicant's creditworthiness, it issued a card for an annual fee that would tend to make the card attractive only to persons who traveled relatively frequently. Thus self-selection as well as the financial eligibility criteria of the issuer combined to produce the result that only persons with relatively high incomes and high time costs were likely to use the card. Thus, having a T&E card signaled to the distant merchant that the holder had the same income and consumption characteristics that induced the merchant to extend open book credit to local customers.

The issuing organization bought the transactional paper thus generated at a discount. Even though by present bank-card standards this discount was relatively large, the relation was worthwhile to the merchant: the system not only enabled the merchant to identify a new group of high-income customers and compete for their patronage but also protected him against default risk, performed billing and collection, and, perhaps most important, eliminated the capital costs of extending credit during the billing cycle.

Because the T&E card was a three-party instrument rather than a four-party instrument, the feature of jointness was present on the demand side but not on the supply side. Again, there was one particular distribution of costs between the merchants and the card holders that would bring their demands for the

transactional service into equilibrium. But the card-issuing organization was a single enterprise; periodic adjustment was within its control, and there was no problem of coordinating two enterprises to determine how to distribute charges between card holders and merchants.

The national T&E cards were not the only three-party transaction cards that appeared during these years. Many major oil companies distributed similar cards, but their merchant base was limited primarily to their distributors. A number of banks also distributed three-party cards. Although these cards were accepted by a more heterogeneous set of merchants, their use was limited to the geographic region to which the banking laws limited the bank's deposit-accepting activity. One of the most successful three-party bank cards was BankAmericard. The Bank of America, enjoying the advantage of a large and populous state with relatively permissive statewide branching laws, was able to reach more card holders and merchants than most other three-party bank-card systems.

Several characteristics of the late 1950s and early 1960s set the stage for the introduction and rapid expansion of the four-party bank credit card. Those were years of relatively rapid growth in real income in the United States. The number of high-income, high-time-cost persons increased rapidly, as did the number who traveled frequently outside their own community. Simultaneously, data processing and electronic communications experienced dramatic technological advance, which enhanced the demand for transactional services and, on the supply side, significantly reduced the costs of maintaining accessible documentation on creditworthiness and of billing and collection.

Moreover, as nominal interest rates began to rise by the late l960s, interest costs became a larger fraction of the total cost of extending consumer credit. The comparative advantage of banks and other financial institutions over all but the very largest of the retail chains became ever more decisive as interest costs predominated in the total cost of performing the retail credit function. Finally, there were scale economies from consolidating one consumer's transaction with a number of merchants into a single statement, a single billing, and a single remittance. All these factors favored substituting bank-card systems for the traditional merchant function of extending retail credit.

The four-party bank credit card was introduced in 1966 in order to obtain for bank-card payment systems a ubiquity that, by reason of our geographically restrictive banking laws, could not be obtained by arty single banking enterprise in its deposit acceptance activities. In that year the Bank of America licensed its "BankAmericard" service mark on a nationwide basis. Licensees were authorized to issue cards bearing the logo, to sign up merchants who would accept the card in the area of the licensee's operation, and to engage other banks as agents to expand the merchant base still further.

At about the same time, under the leadership of the major Chicago banks, the Midwest Bank Card system was established as a joint venture among a number

of banks in the Great Lakes area. Shortly thereafter, the Interbank Card Association was formed as a nonprofit membership organization owned by its card-issuing member banks. Its initial purpose was to provide nationwide interchange facilities to a number of regional systems. Among these local programs was the Western States Bank Card Association which owned the "Master Charge" service mark. In 1969, after that card association had joined InterBank, the Master Charge mark was assigned to InterBank and then licensed to all InterBank members. Thus within three or four years, today's major bank-card systems made their appearance. In 1970 the BankAmericard system changed its structure to that of a membership corporation; in 1977 the name of the national organization changed to "Visa" and exclusive rights to the name "BankAmericard" reverted to the Bank of America.

These organizational changes did not alter the fundamental point that these multibank organizations were from their inception four-party systems having the peculiar economic characteristic previously described. Given the distribution of charges between *P* and *M* that would achieve equilibrium in their demands, it was overwhelmingly improbable that the revenue stream from *M* to *M* bank or from *P* to *P* bank would equal the costs of the subset of activities that a particular bank was required by the technology of the payment system to perform; thus some redistribution of those revenues between *M* bank and *P* bank was likely to be necessary for the payment system to compete effectively with alternative mechanisms.

Hence, half a century after Fed coercion resolved this problem of redistributing revenues in the context of four-party check clearance transactions, the bank-card systems confronted the question how to determine the appropriate magnitude of the necessary transfer payment between *M* bank and *P* bank. It makes no difference when addressing this question in the abstract whether the transfer payment is made by card-issuing banks to merchant banks or by merchant banks to card-issuing banks; I will assume, as recent cost patterns suggest, that income from card holders is too small for the average card-issuing bank to cover its costs, whereas income from merchants is, on average, more than sufficient for merchant banks to cover their costs. As shown in Section I, given the assumption about competitive equilibrium stated there, the magnitude of the deficiency must equal the magnitude of the surplus; I will refer to that magnitude as the optimum transfer fee.

The monopsonistic position of *P* bank—which is determined by the direction of the paper flow and hence would be present even if the transfer fee had to move in the opposite direction—implies that each *P* bank cannot be permitted to announce daily the price at which it will buy paper to be billed to its card holders. If a system involved very few *P* banks and *M* banks, bilateral agreements could be negotiated between each *P* bank and *M* bank, and each agreement could establish for some substantial period of time the magnitude of the transfer fee. This approach has two substantial drawbacks in practice. First, the number of agreements to be negotiated in each time period is equal to the

product of the number or *P* banks and the number of *M* banks; second, and probably more important, there is a significant free-rider problem that increases with the number of participants.

Imagine a card system composed of ten *P* banks that act only as purchaser banks and ten *M* banks that act only as merchant banks. Assume that each *P* bank receives from each *M* bank 1 percent of the aggregate paper flow of the system and has 10 percent of the aggregate card-holder base. Assume, finally, that the optimum transfer fee is 1 percent of the face value of the paper and that this fee amounts to $0.30 per item. Although it is subversive of the system as a whole to demand a higher fee, each individual *P* bank faces a strong temptation to do so—let us assume a 10 percent increase in the transfer fee to 1.1 percent, or $0.33. Any individual *P* bank that so behaves, provided that it is unique in demanding an excessive fee, will increase its fee revenues by about 10 percent but will increase the effective costs confronted by each *M* bank only by 1 percent. Even assuming that the *M* banks immediately pass on this cost differential, the merchant discount would be increased by 1 percent on the paper of all *P* banks, for it is not feasible for the *M* banks to discriminate against paper en route to that particular *P* bank without creating, on the part of all the merchants, an incentive to refuse to honor cards issued by that *P* bank; moreover, any endeavor by all merchants selectively to refuse cards issued by a particular *P* bank (at least outside the context of an on-line electronic system) would substantially increase the transaction costs of all merchants and of all card holders. The utility of the system to all participants would diminish, as would the system's viability in competition with other payment systems.

Similar, although perhaps less immediately dramatic, consequences would follow if either the set of *M* banks or the set of merchants chose to absorb the percent cost increase that flows from *P* bank's 10 percent increase in the transfer fee. Some might drop out of the system entirely because of economic losses: others would alter their behavior in less drastic ways to shift from using the card system to using some other payment systems.

These adverse consequences would eventually reduce the transaction volume of the individual P bank that raised the transfer fee, but the adverse effect would be spread across all P banks. The one P bank would realize 100 percent of the revenue gains from its fee increase but would bear only 10 percent of the adverse consequences. More generally, in a card system involving x number of P banks, anyone bank can exploit the monopsonistic position it enjoys over its own paper and can realize 100 percent of the revenue gains while suffering only a fraction of the adverse consequences, that fraction being l/*x*. Accordingly, it is essential that the participants in a four-party payment system collectively adopt some internal mechanism that prevents individual exploitation of the monopsony power endemic to such systems.

As discussed earlier, banks were prevented from exploiting their monopsonistic power in the checking system initially by collective agreements among

clearinghouse members and later by the Fed's coercive tactics. But the problem was resolved for the checking system without explicit recognition of the problem's characteristics, without any inquiry into the costs of the system, at the apparently arbitrary transfer fee of zero, and largely by government coercion rather than agreement. These all make it unlikely that the resolution was optimum when first made, even less likely that the resolution could have continued to be optimum after the enormous changes in check-processing technology.

Compared to the checking system, the bank credit card system has evolved so far under less government intervention with respect to the transfer fee. Perhaps for that reason, perhaps also because there are many institutions for which items transmitted in their capacity as *M* bank are unequal to items received in their capacity as *P* bank, behavioral characteristics of those payment systems more closely correspond with the behavior implied by the theoretical considerations discussed in Section I.

Before those transfer fee arrangements are examined, two important differences between the checking system and bank-card systems should be noted, differences that significantly affect the cost to the parties. First, under the checking system, *M* bears the risk of default: if funds adequate to cover the check are not on deposit at *P* bank when the instrument arrives for payment, the check is dishonored and charged back through the clearance system against *M*'s account with *M* bank. But under the bank-card system, provided that *M* complies with the prescribed authorization procedures, *P* bank guarantees payment by the card holder and thus bears the risk of default. This shifting of risk under the bank-card system obviously increases *P* bank's cost, enhances *M*'s demand for the system, and increases the amount of discount *M* is willing to pay to *M* bank. Thus, one would expect to observe larger transfer fees from *M* banks to *P* bank than those in the checking system.

The second basic difference between the checking and bank-card systems also has the effect of increasing *P* bank's costs of the bank-card system. Because a check forwarded to *P* bank is debited immediately against funds on deposit. *P* bank incurs only minor float costs. Whatever float costs remain are borne either by *M* bank (if it credits *M* 's account on deposit) or by *M* (if his account with *M* bank is not credited until funds are remitted). Float costs under the bank-card system are borne in different proportions from those under the checking system and are substantially greater. The paper generated by the card holder is not issued against any existing deposit with *P* bank; remittance is made by *P* only at the end of the monthly billing cycle. Unlike the check clearance cycle, which takes only a few days, bank-card items will on average be outstanding on *P* bank's books for two weeks before *P* is sent an accounting statement and for about three and a half weeks before *P*'s remittance is received.

Clearly *P* bank bears the cost of this extended period of float, but the incidence of the corresponding benefit on demand is ambiguous. In comparison with use

of a currency or a check method of payment, *P* is the beneficiary, and his demand for the bank-card system should increase. On the other hand, to the extent that the bank-card system is being used by *P* and *M* in lieu of open-book credit, it is *M* whose float costs have been reduced, and his demand should be enhanced.

Before turning to the messy world of reality, it is useful to ask what one would expect to find there, reasoning from the theoretical joint demand and supply model developed in Section 1. Both *M* and *P* banks will be incurring activity costs, and both will be receiving a revenue stream. Because the revenue stream of each probably will not equal its cost stream, one would expect to observe some side payment that will bring the net revenue stream of each bank, after the side payment, back into the same proportion with respect to its cost stream as the proportion between total revenue and total bank costs. Obviously, any side payment that brings those ratios into equality for the two banks (or sets of banks) has the same effect. Equally obviously, the value of all these ratios will, in competitive equilibrium, equal one.

With these features in mind, one can attempt to derive by arm-chair empiricism a picture of both the demand and the supply sides of the bank-card industry as revealed by present behavior. So far as demand is concerned, there is unmistakable evidence that a positive demand exists on the part of many merchants for bank-card services; and, although the evidence is less clear, there are persuasive reasons to believe that a demand exists also on the card holder side and that it also is positive at prevailing transaction levels. No direct observation of the contours of these demand functions is possible; we catch glimpses of segments of the functions only as demand is revealed by the willingness of merchants and card holders to pay for bank-card services. Thus, in our endeavor to explore demand functions, we are led to examine the charges that banks have historically imposed on merchants and card holders.

Before nominal interest rates skyrocketed in early 1980, the bank-card industry imposed substantially all the costs or bank-card transaction services (as opposed to financing services, a distinction developed hereafter) on merchants. Since each merchant bank is free to negotiate whatever arrangement it can with its own set of merchants, enough variance exists among arrangements to make generalization difficult. Typically, though, merchant discounts have been between 2.25 and 3 percent of total transaction dollars, the discount being higher for merchants who have smaller aggregate dollar volumes or who have smaller average dollar amounts per item. To facilitate discussion I assume where precision is not essential that the typical merchant discount is 2.5 percent.

With exceptions to be discussed later, no charge has been imposed on the card holder. In this context, too, each card-issuing bank is free to negotiate such arrangements as it wishes with its card holders. Before 1980 only a few card-issuing banks had imposed either transaction fees or periodic “membership” fees on their card holders: in the overwhelming preponderance of instances,

banks have been willing to play the role of *P* bank as a competitive gambit to attract the individual demand deposits of its card holder. Until recent regulatory reform permitted banks to pay interest on demand deposits, the value to the card-issuing bank of attracting incremental individual demand deposits on which no interest was or could be paid was a sufficient inducement, at least when coupled with the interchange fee received from the merchant bank, to compensate *P* bank. Thus, although revealed demand plainly exists on the merchant side, it is less clear on the card holder side.

The picture is complicated on the card-holder side by the fact that the bank credit card historically has not been merely a payment mechanism. The card holder has had the option of paying, at the end of a billing cycle, only a minor fraction of the charges incurred during that billing cycle and deferring payment of the preponderant portion of the balance. But if he does "revolve" his account in this way, interest payments become due not only on the balance deferred, but also on each new charge subsequently incurred until the balance is, at the end of some billing cycle, reduced to zero. In short, card holders who revolve their accounts not only pay interest on the deferred balances but lose the advantage, available to those who do not revolve, of about three weeks "free" float on current transactions.

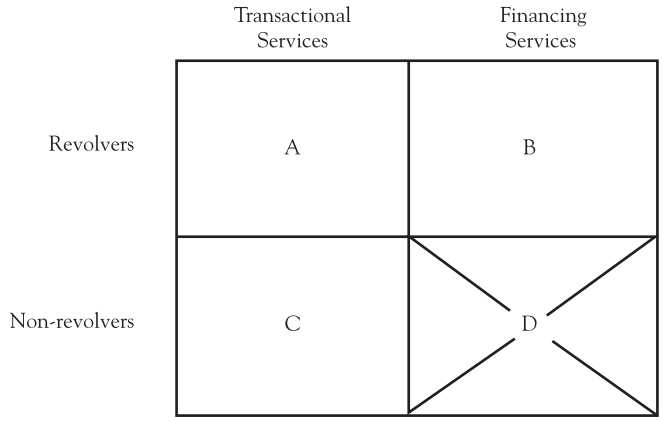

FIGURE 5

Thus the card-issuing bank can be viewed as engaged in two different businesses. It sells a transaction service involving valuable float to those "nonrevolvers" who choose to pay their statement in full at the end of each billing cycle. It also sells a combination transaction service anti consumer finance service to those who use their bank cards as an extended credit mechanism. Because certain activities essential to providing the payment service—receipt of interchange items, posting to individual card holder accounts, billing, collection, posting of credits, bearing the risk of default, etc.—must be performed with respect to revolvers as well as nonrevolvers, complex accounting allocation problems arise.

Several different views of the bank-card industry can be taken. Figure 5 will aid in distinguishing the possible views and the accounting differences that seem to follow from taking one view rather than another. The alternative views present the industry as engaged in only one business or in two different businesses. If the industry is thought to be in two businesses, there are alternate ways of defining those two businesses, if two or more business segments are truly joint (in the sense that one set of services cannot be rendered economically without simultaneously performing the other), it is pointless and potentially misleading to regard them as separate businesses. Equalization of both *P* bank and *M* bank revenue-to-cost ratios throughout all such segments is our theoretical expectation. If jointness in that sense between any two segments is not present, then one should expect to observe an endeavor, first, to engage in cost allocation and revenue allocation as between such disjoint segments and, second, to observe an endeavor to equalize, within each of those segments, the revenue-to-cost ratios of the two sets of banks. The significance of disjointness is that, should the system-wide revenue-to-cost ratio for one such segment consistently fall below the value of one while the ratio for the other segment exceeded one, the former activities would be abandoned as a commercial failure and the latter activities would be continued.

As the matrix in Figure 5 illustrates, the industry provides three distinct services: transaction services to revolvers (cell *A*), financing services to revolvers (cell *B*), and transaction services to nonrevolvers (cell *C*).

One possible "two-business" view separates activities according to the type of service so that the provision of transaction services to revolvers and nonrevolvers is one business, the provision of financing services to revolvers a second. From an accounting standpoint, this view suggests a cost allocation to cell B of (1) the interest cost of the outstanding balances of revolvers; (2) the incremental billing and collection costs, if any, associated with the extended credit function (as distinguished from those associated with the payment mechanism function); and (3) the incremental costs, if any, of risk of default or fraud associated with the extended credit function (as opposed to the payment mechanism function). Under this view, the periodic interest charge to revolvers would be set at a level just sufficient to cover that set of incremental costs. The costs associated with the payment system features of the card, for those transactions engaged in by card holders who regularly took advantage of the extended credit feature and for those transactions by nonrevolvers, would be regarded as payment system costs that would be covered by some other revenue stream, which might consist of the merchant discount or a separately identifiable charge imposed upon all card holders, such as a periodic membership charge or a per-item charge or a per-dollar volume charge. This first view involves the difficult problem of deciding the extent to which bookkeeping costs and risk costs are incrementally associated with the extended credit function.

Alternatively, one could view the industry as being engaged in two businesses but, rather than linking cell *A* with cell *C* and defining cell *B* to be the sepa-

rate business, this second view links cell *A* with cell *B* and defines cell *C* to be a separate business. This view defines the two businesses with reference to card holder payment practices. One business consists of providing transaction and financing services to revolvers; another consists of providing transaction services to nonrevolvers. The implied accounting allocation problem is to allocate each category of banks' activity costs either to revolvers as a group or to nonrevolvers as a group. Under this view, the cost allocation problem is to associate some fraction of total bookkeeping costs and total fraud and default costs with habitual revolvers and the remaining fraction with habitual nonrevolvers. For habitual revolvers, there are three possible revenue sources: periodic interest charges on outstanding balances, the merchant discount, and other card holder charges such as membership or per dollar fees. For nonrevolvers, only the two latter revenue sources are available.

A third view is that the industry engages in a single business. No cost allocation is attempted, three possible revenue sources previously identified are seen as being available to cover all costs.

From a theoretical standpoint it seems clear that cells *B* and *C* are disjoint. One can readily conceive of a bank-card service that did not offer the extended payment feature. Although nothing resembling the financing service that is provided to revolvers would be possible unless a transaction service was being rendered as well, it would be possible for banks to render transaction services without providing financing services. The T&E cards typically do just this. Accordingly, sensible business practice requires that the avoidable costs of the extended credit activity be ascertained and compared with the incremental revenues to assure that a revenue-to-cost ratio of not less than one exists. But if incremental revenues equal or exceed incremental costs, the extended credit function is commercially viable so long as transaction services continue to be provided: no more stringent test—for example, a requirement that total revenue equal or exceed total cost—is appropriate.

C. Modern Developments

Several events since 1980 require significant adjustments by the bank-card industry. Among the most important are the changes introduced by the Depository Institutions Deregulation and Monetary Control Act of 1980.[64] This legislation, and the regulations that implement it, require the Fed to impose cost-based fees on banking institutions to which it renders services, including check-clearing and collection service; authorize the Federal Home Loan Bank Board to render clearing and collection services, again on a cost-

64 Pub. L. No. 96-221, § 1, 94 Stat. 132 (codified at 12 U.S.C. 226 (1980)).

based fee basis, to savings and loan institutions (S&Ls); authorize a significantly broadened scope of activities by S&Ls, including nonbusiness demand deposits (NOW accounts), broadened lending authority, and credit card services; and authorize both banks and S&Ls to pay interest on demand deposits.

The second significant development was the unprecedented escalation in 1980 of nominal interest rates on debt instruments of all maturities and, in particular, the sharp increase in both nominal and real interest rates on short-term paper.

The third development is the decline of usury laws. The Deregulation Act preempts some state usury laws, and some states are moving quickly to raise or remove other usury limits. These several developments comprise a set of diverse and substantial shocks that will require both a short-run and long-run industry adjustment. Some of the short-run adjustments are already quite visible.

The most significant of these recent developments is likely to be the elimination of the prohibition against paying interest on demand deposits. Heretofore, in most urban areas, and some rural areas as well where the structure of the retail banking industry was conducive to rivalry, commercial banks have engaged in vigorous nonprice competition to attract demand deposits. In significant part, this rivalry took the form of a geographic proliferation of retail bank establishments: multiple branches where branching was freely permitted and small independent establishments where it was not. Thus, banks competed for demand deposits by offering potential depositors geographic convenience. Unless one assumes that the interest prohibition had no effect on the industry at all, one must conclude that, at least to some extent, depositors would have preferred interest payments to incremental geographic proximity and that they will now avail themselves of that possibility. Some fraction of existing banking establishments will prove to be uneconomic, but their disappearance will require a long-run adjustment. Bank payment of interest on deposits will be and is being made in the short run. Profitability will be adversely affected until long-run adjustments have occurred.

The other important dimensions on which banks competed for demand deposits included the provision of checking services without the imposition of transaction charges and the "free" provision of collateral services such as safety deposit boxes and bank card issuance. In these dimensions, short-run adjustments are feasible, and the introduction of charges for such collateral services has been widespread. Since 1980 a large fraction of card-issuing banks have imposed either periodic fees or per transaction lees on card holders. Periodic interest charges on the outstanding balances of extended credit users have also been increased by a number of banks. Both of these changes were facilitated by the removal or escalation of usury limits.

It is clear that these various developments have had and will have a substantial effect on the credit card industry. In the past, users of checks have faced

artificially low marginal prices for incremental check transactions. Uncompensated demand balances have yielded adequate bank revenues to cover those costs. The widespread introduction of NOW accounts by S&Ls will erode any remaining supracompetitive profitability associated with demand deposits, increasing pressure to impose transaction charges. And the payment of interest by banks on demand deposits will both add to that effect and alter competitive strategies for attracting demand deposits. The introduction of cost-based fees for federal collection and clearance services also will increase the cost of using checks. All these factors will work together to dissuade the providers of demand deposit services from providing those services without imposing explicit transaction charges. Many depositors who previously received free checking services will now face per item transaction charges, and the level of charges demanded of other depositors will increase. These increases in the marginal cost of using checks will shift out the demand curve for credit cards.

Simultaneously, however, the supply curve for credit card transactions will also be shifting to the right because of the high cost of funds. Not only the height of these functions but also their shapes over the relevant range will undoubtedly change in ways we do not yet know.

As I emphasized in Section I, the shifting cost function under consideration cannot usefully be viewed as reflecting the cost of dealing with card holders; it reflects the joint cost of providing transaction services to both card holders and merchants. Nevertheless, substantially all of the recent price changes are in the charges imposed on card holders rather than in1 the merchant discount.

It would be an astounding coincidence if at the end of this first round of price changes the distribution of charges between card holders and merchants happened to equilibrate the individual demand functions of those two sets of panics so that each set wished to engage in the same number of transactions at the prevailing price. It seems more probable that a lengthy process of adjustment will ensue, during which financial institutions will gravitate by trial and error to some new equilibrium. And it seems equally probable that the new equilibrium will involve either a higher or a lower interchange fee than that presently in effect. As previously explained, the interchange fee for any one card system must be determined collectively by the system's members: any attempt to set that fee bank by bank, to reflect each bank's individual costs (rather than the system's average costs), would invite each bank to free-ride on the others and set inappropriately high fees.

In addition to the present perturbations in the industry, the "debit card" is for the first time being distributed widely. Apparently many institutions in the industry believe that the debit card and the credit card can be combined and embodied in a single set of plastic cards. Transactions using the cards would be subject to the same merchant discount and the same interchange fee notwithstanding that the card-issuing bank would handle the two types of transactions quite differently. This outcome seems most unlikely unless the contractual

terms that have traditionally accompanied the credit card are materially altered. From the standpoint of the card-issuing bank, debit card transactions will be substantially cheaper than credit card transactions, for debit card transactions will not be authorized unless they are for amounts less than the card holder's deposit balance, in which case the default risks are relatively low. Moreover, since the transaction amount is immediately debited against the card holder's deposit balance, the float costs of the debit card are substantially less. These considerations alone seem to dictate quite a different distribution of fees between card holder and merchant and a different interchange fee, as well. In addition to these cost factors, demand factors suggest a similar conclusion. From the card holder's stand-point, the debit card is less attractive than the credit card. The float costs that the bank saves when a debit card is used are precisely the float benefits that the card holder forgoes when he uses a debit card. One would expect therefore that any card holder entitled to use a credit card will always use it rather than a debit card. It follows that the only frequent users of debit cards will be people whose incomes and other indicators of creditworthiness do not enable them to obtain and use credit cards.

The characteristics that distinguish credit card users from debit card users will substantially affect the demand curve of merchants for transactions with these two different types of card holders. The holder of a credit card will continue to be identified as a customer for whose patronage the merchant wishes to compete by extending a free float period: but that will not be true of the holder of a debit card, and one would expect merchants to be unwilling to accept discounts on debit card paper as large as the discounts traditionally accepted on credit card paper. It seems likely, therefore, that the two payment vehicles will have to be differentiated and subjected to different patterns of distributing charges between merchants and card holders and, in all probability, to different interchange fees. Hence I believe that card-issuing institutions will be engaged in not one but two different learning processes in the period immediately ahead; and both processes will be retarded if these institutions are reluctant to recognize the sharply different cost and demand characteristics of the two payment vehicles.

III. CONCLUSION

Four-party payment vehicles such as the check, the credit card, and the debit card are characterized by joint costs and also by interdependent demand on the part of their users, which, despite the antiquity of such mechanisms, neither the economic literature nor the institutions that provide their services have fully recognized. Those characteristics, in my judgment, were an important contributing cause to the controversy over "clearance at par" that troubled the banking industry for more than half a century and was quieted at last only by means of federal coercion and subsidy. A repetition of the same basic controversy in the context of new payment mechanisms—credit cards and debit cards—is likely to occur in the next few years. Because of sharp cost and

demand changes attributable to legislative amendments, because of the effect of inflation on nominal interest rates, and because of governmental responses to inflation that have taken the form of restrictive monetary policies that increase the real interest rates on short-term obligations, those years are likely to be characterized by disequilibrium, confusion, and controversy. In such a period, reliance on governmental intervention to reduce uncertainty is likely to appeal to at least some of the disputants. Such intervention should be resisted.

Once the economic peculiarities that underlie such payment mechanisms are recognized, one can conclude that legal mechanisms already in place are entirely adequate for the task of equilibrating the market. The courts should recognize that collective institutional determination of the interchange fee is both appropriate and desirable. To an unsophisticated observer this collective process of equilibration resembles horizontal price fixing, but, for the reasons set forth in this paper, it should not be so treated. Because of the potential for free-rider behavior, individual establishment of interchange fees will almost certainly produce chaotic results, such as higher fees and instability within card systems.

On the other hand, the fee that is collectively set should not be binding prospectively on any pair of banks within the system. Any pair of banks in the system should be free to negotiate a different bilateral arrangement by higher or lower fees for paper interchanged between them. The collectively determined interchange fee should be merely a guarantee that no card-issuing bank will demand a higher fee on paper presented to it in the absence of such a bilateral arrangement. Of course, the fee should be regarded as binding retroactively for transactions already executed. Sensible administration of section 1 of the Sherman Act, applied in a rule of reason context, is sufficient to arrive at this result.[65]

It seems equally clear that the movement toward a competitive equilibrium requires no other collaborative action between participants in such payment systems. It is entirely compatible with that competitive equilibrium that individual *P* banks compete with respect to the charges imposed on cardholders and *M* banks with respect to the magnitude of the merchant discount.

Although collaboration among competing banks with respect to the interchange fee should be permitted under the antitrust laws, any expansion of the range of cooperative action should be viewed with healthy skepticism. Thus antitrust and banking authorities should be alert to ensure that the number of

65 See Broadcast Music. Inc. v. Columbia Broadcasting Sys., Inc., 44 U.S. 1 (1979); Continental T.V., Inc. v. GTE Sylvania Inc., 433 U.S. 36 (1976). However, the Supreme Court has on occasion failed to recognize the significance of maximum price fixing where the product has joint-demand characteristics. See Albrecht v. Herald Co., 390 U.S. 145 (1968). See also Frank H. Easterbrook, Maximum Price Fixing, 48 U. Chi. L. Rev. 886 (1981).

payment systems is as large as the attainment of scale economies permits. Though unbridled autonomy within a system cannot be attained, unbridled rivalry between a multiplicity of systems should be encouraged.

In this regard it is regrettable that the Antitrust Division did not give a less qualified response in 1975 to Visa's request for a business review letter pertaining to its then-effective prohibition against dual membership. Visa sought advice with respect to a by-law that prohibited any card-issuing bank or any merchant bank in the Visa system from serving simultaneously either as a card-issuing bank or a merchant bank in any other system. In a business review letter dated October 7, 1975, to outside counsel for Visa from the assistant attorney general, the Division gave a blessing so limited and so carefully hedged as to leave unresolved the legal permissibility of an effective prohibition against dual membership. Visa responded by withdrawing all restrictions on dual membership, even the limited restrictions that the Division was willing to condone.[66]

In the last five years dual membership in the Visa system and the MasterCard system has become the rule. This widespread pattern of dual membership predictably created very strong pressures for standardization in equipment, procedures, and format. Intersystem rivalry has not completely disappeared; but the opportunity and incentive for such rivalry, particularly in technological innovation, has greatly diminished. This regrettable loss of competitive structure was avoidable but is now probably irreversible, for political reasons if for no others.

Contributing to this irreversibility is the fact that technological changes in the intervening years have facilitated a great degree of interbank competition within a particular system than appeared possible in 1975. Improvements in communications technology have made it possible for a subgroup of banks within a system, subject to only minimal standardization, to differentiate the financial service they offer or even to deploy a differentiated set of terminals and yet continue to operate within the system network.

Of course the more obvious but nevertheless important forms of interbank competition—for card-holder accounts and for servicing merchants—continue. Although the loss of intersystem rivalry is unfortunate, and although such rivalry should be carefully preserved if a new opportunity, in the form of a new card system, arrives on the scene, the industry appears to be functioning competitively.

---

66 See generally Note, New Directions in Bankcard Competition, 30 Cath. U. L. Rev. 65 (1980).

THE COMPETITIVE EFFECTS OF THE COLLECTIVE SETTING OF INTERCHANGE FEES BY PAYMENT CARD SYSTEMS

Howard H. Chang,
NERA Economic Consulting

David S. Evans,
NERA Economic Consulting

Previously published in Antitrust Bulletin, Vol. 45, Fall 2000.

Abstract

In four-party payment card systems, members use the cooperative to achieve both traditional economies of scale and the network economies associated with balancing the merchant and cardholder sides of a two-sided market. Coordination, fine-tuned with interchange fees, maximizes the total value of the payments service. Alternatives to collective setting of interchange fees, varying from bilateral negotiation to government-regulated cost-based fees, all have serious drawbacks in terms of generating excessive transactions costs, failing to internalize external benefits and costs, and distorting incentives.

Collective interchange fees have proved superior to the alternatives from a variety of perspectives. Fees charged to merchants by issuers in the card associations are historically lower than fees set by three-party proprietary systems. And any restrictions on interchange fees (let alone a ban) would bias regulation in favor of the closed systems, thereby undermining competition between open and closed systems.

Analogies between payment card interchange and check interchange, which operates with a zero interchange fee, are misleading. The Federal Reserve's ability to drive check interchange rates to zero does not demonstrate that a zero-fee system represents a competitive, welfare-maximizing equilibrium.

I. INTRODUCTION

Consumers charged almost $1 trillion in purchases on payment cards in 1998, amounting to around 17% of all expenditures made by consumers for personal consumption that year.[1] Payment cards include credit, debit, and charge cards issued by American Express, Discover, Diners Club, and several membership associations of financial institutions: MasterCard International (MasterCard), Visa U.S.A., Inc. (Visa), and regional ATM networks.[2] In 1998, Visa brand cards had a 48% share of the volume of dollars charged on payment cards, MasterCard had 24%, American Express 17%, the regional ATM networks 5%, Discover 5%, and Diners Club 1%, the regional ATM networks 5%, Discover 5%, and Diners Club 1%.[3] Collectively, payment card associations account for about 72% of all charges made on payment cards.

Consumers can use their payment cards at millions of merchant locations in this country and millions more abroad. For example, Visa cardholders can use their cards for payment at almost 4 million merchant locations in the U.S. and at another 12.7 million locations abroad.[4] American Express, Discover, and members of the MasterCard, Visa, and regional ATM associations enter into contracts with merchants to take their respective cards for payment. Merchants pay a "merchant discount"—usually a percent of the transaction amount—to the payment card entity that processes their transactions. In the case of the payment card associations, the payment card entity that processes the merchant transaction (known as the acquirer) often differs from the payment card entity that issued the card used in the transaction (known as the issuer). The payment card associations have adopted rules that require the acquirer to pay the issuer an "interchange fee." The interchange fee is usually a percent of the transaction amount. For example, MasterCard acquirers pay MasterCard issuers about 1.4% for most retail transactions.[5]

---

* Economists with National Economic Research Associates, Inc. in Chicago, Illinois.

AUTHORS' NOTE: *We thank Richard Schmalensee, Daniel Garcia Swartz, and Larry White for helpful comments and discussions on the topics examined in this article. Some of the research reported in this article was supported by Visa U.S.A., Inc. The views expressed herein, however, are solely those of the authors.*

1 *Table 3: Gross Domestic Product and Related Measures: Level and Change From Preceding Period* (visited June 25, 1999) <http://www.bea.doc.gov/bea/dn/niptbl-d.htm>; THE NILSON REP., No. 689 (April 1999), No. 687 (March 1999).

2 In addition to these types of cards, which can be used at millions of merchants, some businesses (e.g. Sears) issue payment cards that can be used only at establishments affiliated with those businesses.

3 THE NILSON REP., No. 687 (March 1999), No. 689 (April 1999).

4 THE NILSON REP., No. 689 (April 1999), No. 691 (May 1999).

5 Jason Fargo, *The Quest for New Markets*, 11 CREDIT CARD MGMT. 56 (March 1999).

The collective setting of the interchange fee has drawn antitrust scrutiny. In the early 1980s, Visa was sued by National Bancard Corporation (NaBanco).[6] The plaintiff asserted that Visa's interchange fee was a per se illegal agreement or, at a minimum, violated the rule of reason. The Eleventh Circuit rejected both of those challenges in its 1986 decision, noting that the lower court's finding that the interchange fee "is more procompetitive than anticompetitive is supported by substantial and persuasive evidence."[7] Nevertheless, the interchange fee has remained controversial. Dennis Carlton and Alan Frankel have argued that the economic reasoning in *NaBanco* was flawed, and their analysis "leaves open the possibility of an antitrust challenge to interchange fees."[8] Alan Frankel[9] has suggested that the collective setting of the interchange fee is part of a long tradition, dating back at least to the 19th century, of the exercise of market power in payment mechanisms.[10]

In this article, we explain the economic role of interchange fees in payment card associations. We show that collectively set interchange fees cannot generate economic rents for members of these associations and demonstrate that an interchange fee set by government intervention would likely reduce social welfare. In section II, we provide some background on the payment card industry and explain how the interchange fee addresses complex coordination and incentive problems for the associations. In section III, we present an economic model that demonstrates that the zero interchange fee seemingly advocated by Carlton and Frankel is not generally optimal, and that government interventions that result in a zero interchange fee would not generally increase social welfare. In section IV, we discuss several of the concerns that commentators, in particular Alan Frankel, have raised about the collective setting of interchange fees. Furthermore, we show that the structure of the payment card associations preclude either associations or individual members from realizing economic rents through the collective setting of the interchange fee. In section V, we present a brief summary.

---

6 National Bancard Corporation v. Visa U.S.A., Inc., 596 F. Supp. 1231 (S.D. Fla. 1984), *aff'd*, 779 F.2d 592 (11th Cir. 1986), *cert. denied*, 479 U.S. 923 (1986).

7 779 F.2d at 606.

8 Dennis W. Carlton & Alan S. Frankel, *The Antitrust Economics of Credit Card Networks*, 63 ANTITRUST L.J. 643, 661 (1995). *See also* David S. Evans & Richard Schmalensee, *Economic Aspects of Payment Card Systems and Antitrust Policy Towards Joint Ventures*, 63 ANTITRUST L.J. 861 (1995).

9 Alan S. Frankel, *Monopoly and Competition in the Supply and Exchange of Money*, 66 ANTITRUST L.J. 313 (1998).

10 The interchange fee is also a central focus in an antitrust challenge by a group of retailers to the honor-all-cards rule adopted by the Visa and MasterCard associations. *See Second Amended Consolidated Class Action Complaint, In Re* Visa Check/MasterMoney Antitrust Litigation, No. CV 96-5238 (D. E.D.N.Y., Oct. 25, 1996), May 26, 1999. The plaintiff retailers claim that this rule requires them to take the associa tions' debit cards as a condition of taking their credit cards. Although the plaintiff retailers have challenged the honor-all-cards rule on the ground that it is a per se illegal tie, they have claimed that but for the alleged tie the interchange fee on debit cards would be far lower. *See Memorandum of Law in Support of Plaintiffs' Motion for Class Certification, In Re* Visa Check/MasterMoney Antitrust Litigation, at 18, April 15, 1999.

II. BACKGROUND ON THE INDUSTRY

American Express (in the U.S.) and Discover are "closed" systems. They consist of a single entity that issues cards to individuals and processes transactions for merchants. American Express and Discover are for-profit companies. Visa, MasterCard, and many of the regional ATM networks are associations of banks with open membership. These payment card associations authorize transactions and settle accounts between members. They also set rules and engage in other activities such as brand advertising and promotion that assist the members of their respective associations. The associations themselves do not issue cards, nor do they establish the terms and conditions under which cards are issued to individuals. Although the members of these associations are for-profit businesses, MasterCard, Visa, and most of the regional ATM associations operate on a not-for-profit basis.

The payment card business is unique among major industries for several reasons. The demand and supply of plastic cards involves the joint interaction and necessary participation of the systems, cardholders, merchants, issuers, and acquirers.[11] These parties interact organically within a complex system. Actions taken by any one of these parties necessarily affects the other parties. Two striking economic characteristics of the plastic card business create strong interdependencies among these parties. First, the demand for a particular brand of payment card results from a joint decision by people to have and use that card brand and by merchants to take that card brand. Both the cardholder and merchant obtain the benefits from a card transaction.[12] Second, the demand for a particular brand of plastic cards is subject to what economists call network effects or positive-feedback effects. Cardholders value a particular card brand more if more merchants accept it. Merchants value a particular card brand more if more customers have it.[13]

In part A, we explore the business interactions among the parties to payment card transactions. In part B, we explain the significance of interdependent joint demand and network effects for these interactions. In part C, we describe how the associations coordinate cardholder and merchant demand and harness network effects to increase the value of their respective brands. For expositional simplicity, we focus on MasterCard and Visa.

---

11 In the case of the closed systems, the system conducts all issuing and acquiring.

12 William F. Baxter, *Bank Interchange of Transactional Paper: Legal and Economic Perspectives*, 26 J.L. & ECON. 541, 573 (1983); DONALD I. BAKER & ROLAND E. BRANDEL, THE LAW OF ELECTRONIC FUND TRANSFER SYSTEMS ¶¶21.03, 24.01 (1998).

13 BAKER & BRANDEL, *supra* note 12, ¶21.03; Evans & Schmalensee, *supra* note 8, at 887.

A. Visa and MasterCard transactions involve five interdependent parties

Visa is an association of banks that provides a network service based on guaranteeing cardholders that their cards will be accepted everywhere they see the blue-white-and-gold Visa "flag" (or logo), and guaranteeing merchants that they will receive payment for authorized charges on cards with the blue-white-and-gold flag. Visa has over 6600 issuer member banks.[14] These banks have 108 million active cardholders with 249 million cards and approximately 4 million merchant locations that accept Visa cards.[15] Similar statements apply to MasterCard.

The provision of this network service requires the interaction of five different economic actors: issuers, cardholders, acquirers, merchants, and the system.

1. ISSUERS COMPETE FOR CARDHOLDERS

Issuers compete for cardholders by providing multiple card services at various prices. They earn revenues primarily from two sources: (1) the fees they charge cardholders and the finance charges they earn from people who choose to finance some portion of their monthly charges on their credit cards; and (2) interchange fees that are paid by acquirers. Banks compare revenues and costs from plastic cards to determine whether they can expect to earn profits from issuing plastic cards at all, and to determine how many plastic cards to issue and to whom.

2. CARDHOLDERS USE CARDS AT MERCHANTS THAT DISPLAY SYSTEM LOGO

The whole point of having a payment card *system* as opposed to a series of disconnected brands issued by the various banks to their respective customers is that Visa cardholders, for example, can pay for goods and services at any merchant that has chosen to display the blue-white-and-gold Visa flag. In addition to receiving payment services, cardholders can finance payments if they have a credit card with an available credit line or if they have a debit card with overdraft protection or some other credit line attached to their checking account. Issuers compete for cardholders by offering various combinations of fees, services, and incentives. In the case of credit cards, fees consist of finance charges, annual fees, late fees, and other charges; services consist of things like insurance and a cash line for emergencies; and incentives consist primarily of airline miles, credit toward the purchase of gasoline or other products, or charitable contributions. As for debit cards, the issuer usually includes the debit card as part of its checking-account package.

---

14 THE NILSON REP., No. 640 (March 1997).

15 THE NILSON REP., No. 689 (April 1999).

3. ACQUIRERS SIGN UP MERCHANTS FOR THE CARD SYSTEM

Acquirers sign up merchants for the system. They provide guaranteed payment and data-reporting services to these merchants. They earn revenues from charging merchants a merchant discount fee. Acquirers usually have to pay issuers an interchange fee that is determined by the system as discussed below. Thus, ignoring other banking relationships that may exist between the merchant and the acquirer, the interchange fee generally sets a lower bound on the merchant discount charged by acquirers.

4. MERCHANTS HAVE ASSURANCES WHEN THEY ACCEPT CARDS THAT DISPLAY THE SYSTEM LOGO

Just as cardholders know that they can use their Visa card at all merchants who display the Visa flag merchants know that by agreeing to accept cards bearing the Visa blue-white-and-gold flag, they are guaranteed full payment for authorized charges on any Visa-brand cards. A Visa merchant also obtains access to a group of cardholders, some of whom frequent the merchant only if it accepts Visa cards or may buy more at the merchant than they would if the merchant did not accept Visa cards. Merchants decide whether to take Visa cards based on a comparison of the benefits and the costs.

5. THE PAYMENT CARD SYSTEMS COORDINATE CONSUMERS, MERCHANTS, AND BANKS

The Visa and MasterCard systems engage in several activities that affect the four parties above. System management promotes the system brand through advertising and other efforts; establishes the interchange fee that acquirers pay issuers; and operates a computer network for authorizing and settling payments between merchants, acquirers, and issuers.

It is also important to understand what Visa and MasterCard *do not* do. They do not issue cards to consumers. They do not set any cardholder or merchant fees associated with the systems' cards (annual fees, transaction fees, merchant discounts, etc.). They also do not set the interest rate on the credit lines associated with either credit or debit cards. Card fees and interest rates are set individually by the several thousand member banks.[16] There is a wide variation in the fees, interest rates, and features offered by issuers to cardholders.[17]

---

16 THE NILSON REP., No. 640 (March 1997).

17 DAVID S. EVANS & RICHARD SCHMALENSEE, PAYING WITH PLASTIC: THE DIGITAL REVOLUTION IN BUYING AND BORROWING ch. 7 (1999).

B. Payment cards are characterized by joint demand and network effects

The structure that we have just described makes the payment card industry quite unusual. There are two levels of competition: (1) between card brands—sometimes referred to as system competition; and (2) within card brands in the case of the open systems—sometimes called intrasystem competition.[18] Competition is intense among issuers and acquirers.[19] Furthermore, in the case of the open systems there are three intermediaries between the merchant and the cardholder for any transaction—the acquirer and the issuer with the system standing between the two.

It is not just this structure, however, that makes the plastic card business unique. This business has two fundamental economic characteristics that, taken together, distinguish it from any other business we can think of. First, the product is purchased jointly by merchants and cardholders. There is "joint demand." Second, there are strong "positive-feedback effects." In other words, the value of the product to cardholders is higher if there are more merchants that take the card, and the value of the product to merchants is higher if more cardholders use the card.

1. CONSUMERS AND MERCHANTS DEMAND CARD SERVICES JOINTLY

The demand for a particular brand of plastic cards results from a joint decision by (a) people to have and use that card brand, and (b) merchants to accept that card brand. Customers cannot use a card brand at merchants that do not take that card brand; merchants cannot accept cards that customers do not present. The total volume of transactions placed on a particular brand of payment card is therefore determined through the separate, but obviously interdependent, decisions of customers and merchants. In economic terms, both of their demands must be met for any transaction to take place. That is what we mean by "interdependent joint demand."[20]

The fact that the merchant and the customer jointly demand card services and jointly benefit from their consumption of these card services has an economic implication. The card system can potentially recover its costs and earn a profit from some combination of both the merchant and the cardholder. American Express, for example, recovers the costs of operating its system from merchants and cardholders. It charges merchants a merchant discount and charges cardholders an annual fee for its flagship charge card product. If the cost of operating

18 In addition, there is competition between members of the open systems and the proprietary systems.

19 EVANS & SCHMALENSEE, *supra* note 17, at chs. 6 & 9.

20 The demand is joint because it arises from two parties. It is interdependent because one demand cannot be met without the other.

charge card product. If the cost of operating the American Express system increases, it can attempt to recover that cost from both beneficiaries of its services subject to some of the constraints we discuss below. The same is true for the Visa and MasterCard systems except that there is the added complexity that two different entities are typically involved on the issuing and acquiring sides of a given transaction.

2. NETWORK EFFECTS LEAD TO THE "CHICKEN-AND-EGG" PROBLEM

Cardholder and merchant demands are interdependent in another important way. The demand by individuals for a particular card brand is an increasing function of the number of merchants that take that card brand. The function increases because individuals have a higher degree of certainty that they will be able to pay with their cards when they go shopping and have a greater variety of merchants to choose from that take the card they want to use. Likewise, the demand by merchants for a particular card brand is an increasing function of the number of customers who have and want to use that brand. That is because merchants expect greater additional sales and provide greater customer service if more customers have and want to use the card brand. Through positive feedback, the value of the card brand to each cardholder and merchant increases with the number of people and merchants that choose to participate in the card network. Economists sometimes call these positive-feedback effects network effects.[21] For example, the telephone system is said to have network effects because the value of the system is greater the more people each subscriber can call.

Network effects, in turn, give rise to the so-called chicken-and-egg problem for card systems. Consider the development of a card system. At the beginning, the system has no cardholders and no merchants. The system will have no luck persuading consumers to take its card if it cannot convince them that merchants will accept its card. And it will have no luck persuading merchants to accept its card if no consumers have and use its card. Solving this conundrum has been the signature issue for incipient card systems.[22]

---

21 Michael L. Katz & Carl Shapiro, *Systems Competition and Network Effects*, 8 J. ECON. PERSP. 93 (Spring 1994).

22 Many efforts to establish payment cards have, indeed, failed because the entrepreneurs could not persuade enough consumers or merchants to take their cards. For example, during the 1950s many small banks in the Northeast established payment card programs. The vast majority of these programs failed. A major reason these banks failed was that they found it difficult to establish a sufficient merchant base to appeal to cardholders because that involved soliciting merchants beyond the immediate vicinity of the bank's office. *See* GAVIN SPOFFORD & ROBERT H. GRANT, THE FEDERAL HOME LOAN BANK BOARD, A HISTORY OF BANK CREDIT CARDS 8-12 (1975).

At its inception, for example, American Express bought up existing portfolios of restaurant and hotel charge cards.[23] These acquisitions enabled American Express to solicit merchants to accept its new card by assuring them that there would be a substantial cardholder base right away. When American Express decided to enter the credit card business with Optima in 1987, it used its existing base of merchants to also provide acceptance for Optima and solicited cardholders from its existing charge cardholder base.[24] Similarly, in 1959 Bank of America was able to assure merchants that it would have a cardholder base for its BankAmericards because it issued cards to its banking customers (in those days, it was permissible to issue unsolicited credit and charge cards to consumers).[25]

C. The payment card systems provide rules and balance cardholder and merchant demand

Positive-feedback effects and interdependent joint demand create difficult coordination problems. Not surprisingly, the various card systems have developed similar business procedures to deal with these coordination problems and to harvest network effects.

1. UNIVERSAL ACCEPTANCE OF CARD BRANDS IS COMMON AMONG CARD SYSTEMS

Card *systems* arose because individuals wanted cards they could use widely and merchants wanted cards that many people had. That is why the Balkanized local credit card programs of the late 1950s failed miserably. And that is why every successful card system, as far as we have been able to determine, requires merchants who display a card system's logo (or flag) to take all cards bearing the logo of that system. Visa's honor-all-cards rule has analogues for MasterCard, American Express, Discover, Diners Club, and the point-of-sale (POS) debit transactions on the regional ATM systems.[26] These rules help ensure that cardholders are certain that merchants that display the card

23 *See* PETER Z. GROSSMAN, AMERICAN EXPRESS: THE UNOFFICIAL HISTORY OF THE PEOPLE WHO BUILT THE GREAT FINANCIAL EMPIRE 283 (1987); JON FREEDMAN & JOHN MEEHAN, HOUSE OF CARDS: INSIDE THE TROUBLED EMPIRE OF AMERICAN EXPRESS 53 (1992).

24 Eric N. Berg, *13.5% American Express Rate*, N.Y. TIMES, March 11, 1987, at D1; and THE NILSON REP., NO. 401 (April 1987).

25 JOSEPH NOCERA, A PIECE OF THE ACTION: HOW THE MIDDLE CLASS JOINED THE MONEY CLASS 26-27 (1994).

26 American Express Co., *Form 10-K*, Fiscal Year Ended December 31, 1987; THE NILSON REP., No. 420 (January 1988); *The Discover Card Soon Will Get Some Siblings*, CREDIT CARD NEWS, April 15, 1995, at 1; *Dean Witter Ups the Ante with a Two-Tiered Credit Line*, CREDIT CARD NEWS, September 15, 1995, at 1; Peter Lucas, *Discover's New Chemistry*, CREDIT CARD MGMT., March 1996, at 50.

system's logo will in fact accept payment with that card. They enable systems to offer a national (indeed, international) "currency," instead of many different mediums of exchange, none of which is generally held or widely accepted.

Most card systems[27] also prohibit merchants from charging more for customers that use their cards for a transaction than for customers who use other brands of cards or cash or checks.[28] This rule has at least two important effects. First, it helps the enforcement of the honor-all-cards rule because otherwise a merchant effectively could decline any card by imposing a sufficiently high surcharge on that transaction. And second, it provides cardholders a clear understanding of the costs of their cards. A cardholder agrees to pay whatever annual fees and finance charges imposed by the issuer but knows she will not have to pay any additional surcharge at the merchant.

2. INTERCHANGE FEES COORDINATE THE MERCHANT AND CARDHOLDER SIDES OF THE BUSINESS

The ability to coordinate the cardholder and merchant sides of the business has been essential to the success of payment card systems. For open systems, interchange fees have been an essential tool of such coordination. All payment card systems collect revenues from both sides of the system and thereby balance the incentive to increase the base of merchants that take the card and the base of consumers that have and use the card.[29] The merchant discount is one of the more visible (but certainly not the only) mechanisms that systems use to accomplish this balancing act. For example, American Express cards carry the highest merchant discount—about 2.7% on average[30]—of all major systems. American Express has chosen to have a smaller merchant base and to obtain a relatively higher fraction of its revenues from merchants than have other card systems. American Express receives about two-thirds of its card-

27 Visa, MasterCard, American Express, Discover, Diners Club and some of the ATM systems have such a rule. Most ATM systems, however, permit surcharging both on ATM and retail transactions. *See Interlink Opens the Gates Wider for a Debit POS Surcharge Rush*, DEBIT CARD NEWS, November 18, 1996, at 2; *MAC Fires a Tentative Shot Across Visa's Debit Bow*, BANK NETWORK NEWS, September 11, 1998.

28 Discounting for cash or checks is often permitted.

29 Of course, it would be possible conceptually to have a card system in which merchants are charged nothing for card transactions. The system (and its members) would earn their profits entirely from cardholders. By giving the payment service away for free to merchants, the system could increase merchant acceptance and therefore the value of the card to cardholders. Of course, the system would charge cardholders more. Likewise, it would be possible conceptually to have a card system in which cardholders are charged nothing, or in which the issuer provides below-cost financing to credit cardholders who revolve. By giving the service away to consumers, or even paying consumers to take it, the system could increase consumer acceptance and use and therefore the value of the card to merchants. Of course, the system would charge merchants more.

30 AMERICAN EXPRESS CO., 1998 ANNUAL REPORT (1999).

related revenues from merchants,[31] compared to about one-quarter for Discover.[32]

To balance the merchant and cardholder sides of the business, the card associations must solve a particular problem. Unlike American Express and Discover, which deal with cardholders and merchants directly (i.e., they are both the issuer and the acquirer in every transaction), the card associations have a decentralized system in which the issuer and the acquirer are often different. Thus, there are two independent entities involved that need revenues from the provision of Visa (MasterCard) card services if they are to remain in business. The interchange fee determines the extent to which the issuer and the acquirer share the joint costs and joint benefits arising from the decision by a cardholder to use her card at a merchant that takes that card brand. The interchange fee is the major component of acquirers' costs of servicing merchants.

The Visa (MasterCard) system uses the interchange fee to encourage member banks to issue cards. In the case of credit cards, the interchange fee is the major source of revenue for cards that are issued to individuals who do not incur finance charges. Approximately 40% of consumers are pure "transactors"[33]—people who almost always pay their bills on time and therefore do not avail themselves of the financing alternative. Moreover, because revolvers will often have a payment card on which there are no balances, about 65% of all transactions are made on cards that do not have revolving balances.[34] In the case of debit cards, the interchange fee is currently the main source of revenue for all cards.

For all of the reasons discussed, in setting the interchange fee, the Visa and MasterCard systems have to weigh the effects of the interchange fee on increasing card issuance on the one hand and reducing merchant acceptance on the other hand. If the interchange fee is "too high," merchant acceptance will fall and the value of the card brand to cardholders will fall too. In addition, rather than dropping a card brand entirely, a merchant might undertake efforts to steer consumers away from using that brand.[35] If the interchange fee is "too low," issuers will raise the fees to cardholders and thereby reduce the value of the card brand to merchants. Therefore, the interchange fee is the

31 *See id.*

32 MORGAN STANLEY DEAN WITTER & CO., 1998 ANNUAL REPORT (1999).

33 Calculations based on data from Visa U.S.A.

34 Calculations based on data from Visa U.S.A.

35 In addition, the ability of merchants to promote their proprietary store card programs can be an important factor in affecting interchange fees.

major tool used by these associations to manage the chicken-and-egg problem and coordinate the joint demand by merchants and cardholders.[36]

3. THE NECESSITY OF INTERCHANGE FEES

Some commentators have argued that the collective setting of interchange fees either is[37] or may be anticompetitive.[38] While we consider the merits of this argument later, it is useful at this point to consider the consequences of the obvious potential remedy to this supposed competitive problem: a prohibition on the setting on *any* interchange fee including zero.[39] To examine the consequences of this remedy we must hypothesize about how payment card obligations would be redeemed in a world without interchange fees. Let us start with the real world. Suppose you go into a department store and spend $100 on a really expensive tie. You pay with your MasterCard, the merchant gets the transaction authorized, and you sign a slip. What is that slip worth? With an interchange fee of say 1.5%, the merchant's acquirer knows that it has to pay the issuer interchange of $1.50 on this transaction, but it also knows that it has the right to collect $98.50 from the issuer for this transaction. Under this system, the acquirer has both an obligation to pay interchange and also a right to collect for the transaction (less interchange). The issuer has a corresponding right to collect interchange and an obligation to pay the acquirer for the transaction.

Now, suppose that there is no interchange fee. If MasterCard, for example, were not permitted to specify an interchange fee, then it is no longer clear what a MasterCard transaction represents. What is the department store's acquirer to do with the credit card slip for $100? The acquirer could try to collect $100 from the issuer but the issuer no longer has an obligation to pay it. This suggests the acquirer will not permit the department store to accept the transaction unless it has already worked out a deal with the issuer in advance. To replicate the uniform acceptance that MasterCards currently enjoy, the thousands of issuers and acquirers would all have to reach millions of independent agreements to accept each others' cards. The agreements might even need to be merchant specific, especially for the larger merchants.

---

36 Richard Schmalensee, Payment Systems and Interchange Fees (May 1999) (mimeo). *See also* Baxter, *supra* note 12.

37 National Bancard Corporations v. Visa U.S.A., Inc., 596 F. Supp. 1231 (S.D. Fla. 1984), *aff'd*, 779 F.2d 592 (11th Cir. 1986), *cert. denied*, 479 U.S. 923 (1986).

38 Carlton & Frankel, *supra* note 8, at 661; Frankel, *supra* note 9.

39 In fact, this was one of the remedies sought by NaBanco. *See National Bancard Corporation*, 596 F. Supp. at 1241.

There are clearly substantial costs associated with the bargaining associated with these agreements. Each agreement would need to specify not only the interchange fee to be paid but also the many other details governed by the MasterCard system. For example, issuers currently bear the risk of fraudulent use or of nonpayment by cardholders. There is no reason that would be the case under the independent agreements. Consider a merchant that believes it has a particularly low incidence of fraudulent use. That merchant might want issuers to offer it a lower interchange fee because the risk of fraud on its MasterCard transactions is lower than at other merchants. If the merchant has better information regarding its level of risk than issuers have, then the merchant may prefer to "self-insure" for fraudulent use in return for a lower interchange fee.

The most significant bargaining cost is the risk that some merchants or acquirers do not reach agreements with some issuers. As in all negotiations, there is a risk that parties will not make a deal.[40] In this situation, however, there are specific reasons why it might not be in the interest of some merchants or acquirers to reach agreement with some issuers and vice versa. For example, merchants are likely to differ in terms of the merchant discount they are willing to pay on a MasterCard transaction. Some merchants might only want MasterCard transactions as long as the associated interchange fee is especially low. A merchant's reservation interchange fee might be so low that it would only accept transactions from three-quarters or one-half of all MasterCard issuers. (Even if merchants were not allowed to bargain individually, it is likely that niche acquirers would offer merchants this same deal.) While this strategy is profitable for this particular merchant, its selective acceptance of MasterCards eliminates the possibility of universal acceptance and harms other parties, including other merchants, cardholders, issuers, and acquirers.

Alternatively, consider an issuer that might seek to be a low-cost issuer. An issuer could decide to only authorize transactions from acquirers that were willing to pay a high interchange fee. With higher interchange fees, the issuer might be able to offer larger rebates on transaction volume or other incentives. This comes, of course, at the cost of a diminished merchant base where this particular card is accepted. With some uncertainty, cardholders that had a card declined at a merchant might not know whether it was because the merchant wanted a low interchange fee or the issuer wanted a high one. Even if cardholders know that it is their issuer that is "responsible" for the diminished merchant acceptance, they might still be willing to hold the card and use it whenever possible. Again, this practice of selective authorization might benefit some issuers and some cardholders, but necessarily comes at the expense of the other parties. Acquirers, for example, could no longer offer merchants the ability to accept all MasterCards.

---

40 *See generally* HOWARD RAIFFA, THE ART AND SCIENCE OF NEGOTIATION (1982).

It thus seems likely that prohibiting any interchange fee (even zero) would have severe consequences for the card associations. The potential "solution" that has been suggested is to mandate a zero fee. However, as we discuss in detail in the rest of this article, there are strong reasons to question whether such a solution would benefit consumers.

III. A MODEL OF INTERCHANGE FEE DETERMINATION

In this section, we discuss a simple version of an economic model of interchange fees based on work by Richard Schmalensee. His model analyzes the role of interchange fees in balancing the interests of the issuing and acquiring sides of payment card systems.[41] His model is based on two important assumptions. First, the value of the payment system to issuers is affected by acquirers' actions and vice versa. As we discussed in section II, this interdependence is important in payment card systems. The second assumption is that issuers and/or acquirers do not face perfect competition. While competition does appear to be intense among issuers and acquirers, it is difficult to model interchange fee determination without assuming some slight degree of imperfection in at least one side of the system.[42]

We use a simple mathematical model to simplify the explanation of interchange fee setting. The total volume of transactions made on cards (Q^T) depends on extent to which merchants accept cards (Q^a) and the extent to which consumers have and use cards (Q^i). Card acceptance among merchants results from decisions made by acquirers while card acceptance among consumers results from decisions made by issuers. There are synergies between merchant and consumer acceptance: an increase in merchant acceptance, for example, will lead to an increase in issuers' output. A simple mathematical way to state these relationships is:

$$Q^T = Q^a \text{ (acquirers' actions)} \left(Q^i\right) \text{ (issuers' actions).}$$

We discuss the case where there is only one issuer and one acquirer in the association, but the qualitative results we discuss hold even when there are multiple issuers and acquirers in the system.

As in any economic analysis of decisions, we need to consider how the profits of the decisionmakers are determined. In the case of the acquirer, profits equal the total volume of transactions times the net revenues per transaction.

41 Schmalensee, *supra* note 36.

42 Evans & Schmalensee, *supra* note 8, at 899-901, show that in a perfectly competitive world without any frictions, *any* interchange fee is consistent with a zero-profit market equilibrium. In such a world, if merchants cannot give a discount for cash or impose a surcharge for credit purchases, a single interchange fee, determined by costs and equal to zero only by chance, is consistent with market equilibrium.

Net revenues per transaction equal the price the acquirer receives from the merchant less the cost per transaction for the merchant less (or plus) the interchange fee that the acquirer has to pay to (receives from) the issuer. The equation is:

$$\pi^a = \left[Q^a\left(P^a\right)Q^i\right]\left(P^a - C^a - T\right),$$

where P^a is the price the acquirer charges merchants, C^a is the acquirer's constant per-transaction cost, and T is the per-transaction interchange fee acquirer pays (T is positive when interchange flows from acquirers to issuers). Similarly, the issuer's profit is

$$\pi^i = \left[Q^i\left(P^i\right)Q^a\right]\left(P^i - C^i + T\right),$$

where P^i is the price the issuer charges cardholders, C^i is the issuer's constant per-transaction cost, and T is the interchange fee the issuer receives.

To figure out the association's solution to interchange fee setting, we assume that issuer and acquirer independently choose their prices to maximize their respective profit functions and that the association sets the interchange fee T, to maximize total system profits V where[43]

$$V = \pi^a + \pi^i.$$

Finally, we need to specify how the demand by merchants and the demand by consumers are demand. To make the math as simple as possible, we assume that the demand curves depend only on prices in a simple linear way:

$$Q^a = A^a - B^a P^a$$

$$Q^i = A^i - B^i P^i.$$

Under these simple assumptions, it is possible to calculate the optimal interchange fee:

$$T^* = \frac{1}{2}\left[\left(\frac{A^a}{B^a} - \frac{A^i}{B^i}\right) + \left(C^i - C^a\right)\right].$$

43 Technically, we assume that the association plays a two-stage game. In the second stage, the issuer and acquirer maximize profits given the interchange fee. In the first stage, the association determines the interchange fee to maximize the second-stage profits. Schmalensee also considers a more general objective function for the system that weights acquirer and issuer profits differently.

There are two components to the optimal (profit-maximizing) interchange fee: one demand related, the other cost related.[44] The first quantity in parentheses within the brackets depends on differences in the demand facing issuers and acquirers. The second quantity in parentheses depends on differences in the unit costs of the issuer and acquirer.

Several key conclusions follow from this result.

(1) The profit-maximizing interchange fee would be zero only by chance. This could occur if demand and costs conditions for the acquirer and issuer are perfectly symmetric $\left(A^i = A^a,\ B^i = B^a, \text{ and } C^i = C^a\right)$, then the optimal interchange fee would be zero. A zero interchange fee could also result if the demand-related component had the same magnitude as but the opposite sign of the cost-related component.

(2) The system will tend to give subsidies to the side of the system in which these subsidies will have the greatest effect on increasing total system output.[45] To see this, suppose the unit costs of issuers and acquirers are the same, so that the second term is zero. The elasticity of demand depends on the ratio A/B; the higher this ratio the lower the elasticity of demand at any given price. This means that if the issuer's demand is more elastic $\left(A^a / B^a > A^i / B^i\right)$, then optimal interchange will be positive, meaning it flows from the acquirer to the issuer; if merchant demand is more elastic, then optimal interchange will be negative, meaning it flows from the issuer to the acquirer. If consumer demand is more elastic than merchant demand, as is likely, then consumer demand increases relatively more than merchant demand would have increased with the same subsidy.

(3) The associations have an incentive to set interchange so that it flows from the low-cost side to the high-cost side of the system, thus increasing total system demand. To see this, suppose that demand conditions for the issuer and acquirer are symmetric, so that the first term is zero. If acquirer cost is greater than issuer cost, the interchange fee is negative and will flow to the acquirer. Conversely, if issuer cost is greater than acquirer cost, the interchange fee is positive and will flow to the issuer. Issuers bear the risk of fraud and nonpayment. Consequently, issuing is generally regarded as the high-cost side given current risk allocations by payment card associations.

---

44 The profit-maximizing interchange fee is roughly related to the output-maximizing interchange fee. If markups are the same on the issuing and acquiring sides, the two coincide. This is unlike the situation where, for example, a monopolist's price can differ substantially from the socially optimal price. Here, the interchange fee equilibrates the two sides of the system but is not a final "price" that can be used to extract rents.

45 It is also worth noting that the interchange fee cannot solve a form of the "double marginalization" problem that exists in this setup. The acquirer, for example, in considering whether to decrease prices does not consider the resulting increase in profits accruing to the issuer.

(4) The interchange fee is not a "price" in the usual sense of that word—it is not a fee paid by the end users of a card transaction. It is better thought of as a transfer payment between the two sides of the systems. This transfer payment is used to harness the network effects in the system by stimulating the side of the system that is likely to result in the greatest increase in demand; increases in demand in the side of the system that receives interchange fees then stimulates demand on the side of the system that pays interchange fees.

Although this model of interchange fees is based on some special assumptions, we believe it captures the essential features of interchange fees in payment systems, and does a good job of explaining the interchange fees observed in practice. Consider, for example, the setting of interchange fees by Visa. Visa uses many factors in deciding on interchange fees. First, it relies on accounting studies of the respective costs incurred by issuers and acquirers and, as a general matter, uses interchange revenues from the low-cost side of the system (acquiring) to lower the costs for the high-cost side of the system (issuing), thus stimulating demand on the high-cost side. In fact, over 90% of the costs associated with the payment function on credit cards is incurred by issuers.[46] Visa is also apparently mindful of demand elasticities in setting interchange fees. In order to encourage supermarkets to accept Visa, for example, it set a lower interchange fee for supermarkets.[47]

The model discussed in this section suggests that card associations can use interchange fees to increase output but does not explicitly consider social welfare. Recent work by Jean-Charles Rochet and Jean Tirole[48] compares the privately optimal fee for the association with the socially optimal fee. They find that the socially optimal interchange fee is generally non-zero. The intuition for this is straightforward. Suppose that merchants and cardholders both receive benefits from card transactions. If there is a zero interchange fee, the price charged by issuers depends only on their costs. Suppose there is perfect competition so that they price at cost. In that case, only consumers that value a card more than the issuers' costs would take a card. This is socially inefficient, however, if there are net benefits for merchants from card transactions. The socially optimal incentives could be provided to potential cardholders by using the interchange fee to subsidize the costs to issuers of providing cardholder services. Rochet and Tirole find in their model that under certain conditions the privately optimal fee for the association will be the same as the socially optimal interchange fee. Under other conditions, the privately optimal interchange will lead to issuers oversupplying card services.

---

46 Data from Visa U.S.A.

47 Visa also uses interchange fees for more general system objectives that do not fit directly into the framework of the model discussed in this section. For example, Visa set differential rates that provided incentives for merchants to install electronic terminals in the 1980s, thus enhancing overall system efficiency.

48 Jean-Charles Rochet & Jean Tirole, Cooperation Among Competitors: The Economics of Credit Card Associations (March 1, 1999) (mimeo).

But even in that case, the socially optimal interchange fee is still non-zero, so that mandating a zero interchange fee or eliminating the no-discrimination rule has ambiguous implications for social welfare.

IV. CRITICISMS OF INTERCHANGE FEES

Nevertheless, interchange fees are controversial in part because they are set collectively by associations of competitors. In perhaps the most comprehensive attack on interchange fees, Alan Frankel argues that interchange fees for modern payment card systems are part of a long historical line of vertical price restrictions that have reflected the exercise of market power.

> If market power in some new form of money derives from an entrepreneurial endeavor by an innovative firm, then antitrust policy probably will, and should, have little to say about it. But when new payment systems require the cooperation of large segments of the banking industry, it naturally gives rise to the concern that those banks will enact systems and rules that are not necessary to the success of the payment system, but that result in a significant reduction in the benefits that will flow to the public from the new technology. We must think long and hard before agreeing to give large associations of financial institutions the right to impose a tax on the entire retail economy, on the basis of vague and unsupported theories backed up with an appeal to a history that, on closer examination, reveals centuries of monopolization.[49]

His thesis relies on Gresham's law that "bad money drives out the good" and on the observation that "the price paid by a consumer for a product does not vary with modest differences in the costs imposed on the merchant by the customer's choice of brands or payment methods"—a phenomenon he calls price coherence. Price coherence results from transactions costs that deter merchants from imposing surcharges or offering discounts on particular types of payment methods.[50] Interchange fees shift costs from consumers to merchants. As a result of price coherence, merchants cannot impose the interchange fees on those consumers that cause the fees to be incurred. Consumers who use payment methods that carry interchange fees are subsidized through taxes that are imposed on consumers who use payment methods that do not carry interchange fees. Bad (interchange fee-based) money drives out good (no interchange fee-based) money.

---

49 Frankel, *supra* note 9, at 361.

50 *Id.* at 316-17.

A. Theoretical analysis

Frankel presents three key propositions. (1) Price coherence makes it harder for merchants to substitute away from card transactions (given that they are going to accept cards at all) and thereby reduces the elasticity of demand faced by acquirers. (2) Acting collectively, the members of the associations shift costs onto merchants. This is an exercise of market power. (3) Welfare would be enhanced by requiring interchange fees to be zero. We can use the model in the previous section to evaluate these assertions.

1. PRICE COHERENCE AND THE OPTIMAL INTERCHANGE FEE

The optimal interchange fee is determined by the elasticity of merchant demand, the elasticity of issuer demand, and the costs incurred by acquirers and issuers. Although it is possible that price coherence reduces the elasticity of merchant demand (which is inversely related to A^a / B^a), it is unlikely that this would have a material effect on the interchange fee. To see this, suppose that price coherence decreases the slope coefficient in the linear demand equation for merchants by dB; that would make demand less elastic. Then the increase in the interchange fee as a result of this change, dropping superscripts, equals $AdB/\left[2B(B-dB)\right]$, which is on the order of A/B^2. This is only a second-order effect relative to the size of the optimal interchange fee for the system. Suppose, for example, that $A=10, B=5,$ and $dB=0.5$. The effect of the acquirer's demand elasticity to the interchange fee is 1.0 before and 1.1 after the dB change.

2. INTERCHANGE FEES AND MARKET POWER

Open and closed card systems set merchant discounts and card fees to maximize profits. Merchant discounts and card fees are set to take into account the interdependencies between consumer and merchant demand. The fundamental economics is the same. The difference lies in what we are able to observe in practice. In the case of the closed systems, we observe the merchant discount and card fees. We do not observe an interchange fee because there is no need to specify an explicit transfer payment between the acquiring and issuing sides of the business. In the case of the open systems, we observe the merchant discount, card fees, and the interchange fee. There has to be an explicit transfer between acquirers and issuers in the case of the open systems. The interchange fee is a result of the economics of the open systems, not of their market shares or ability to exercise market power. For example, suppose that there were ten equal size systems of which nine were closed and one was open. The open system would charge an interchange fee.

Another way to see that the interchange fee does not reflect the exercise of market power, in the usual sense of that term, is to consider the effect of interchange fees on the association members' profits. For Visa and MasterCard,

there is intense competition among issuers and acquirers. Consequently, a large interchange fee will not have much effect in converting consumer surplus into profit because profits tend to be competed away. Suppose for the sake of argument that the interchange fee is set "too high" by whatever standard, so that "too much" interchange flows to the issuing side of the business. Issuers will then compete by offering consumers lower prices or other incentives, so that any "rents" from interchange will not be retained by issuers.

3. CONSUMER WELFARE AND "ZERO" INTERCHANGE FEES

Frankel's discussion of price coherence is based on the existence of transactions costs. Transactions costs prevent the economy from reaching the social optimum that would exist in the absence of transactions costs. Making Pareto improvements in an economy with transactions costs, however, is problematic.[51] Consider how the introduction of transactions costs based on price coherence would affect the determination of the interchange fee in the model discussed in the previous section.

First, it is not at all clear that transactions costs would have a substantive effect on the interchange fee. It is wrong to assume that merchants would impose a surcharge equal to the interchange fee in the absence of price coherence. The marginal cost of transactions to the merchant for cash and checks is greater than zero.[52] Whether or not those costs are less than the merchant discount is not always clear. In any case, the difference may well be less than the interchange fee.

Second, there is no empirical basis for determining what the interchange fee would be but for price coherence—there is certainly, as we have said, no presumption that but for price coherence it would be zero. Consequently, although infinitesimal reductions in the interchange fee theoretically might increase social welfare, there is no operational method for determining the "right" interchange fee.

Third, price coherence is not the only "transaction cost" or "market imperfection" in the payment card system. On the consumer side there are the well-known problems of moral hazard and adverse selection and inefficiencies brought on by bankruptcy laws. If one were serious about devising a government intervention that could make Pareto improvements in the payment card industry, one would have to consider all of these transactions costs, and not just price coherence.

---

51 *See generally* ANTHONY B. ATKINSON & JOSEPH E. STIGLITZ, LECTURES ON PUBLIC ECONOMICS §§ 12-14 (1980).

52 FOOD MARKETING INSTITUTE, EPS COSTS: A RETAILER'S GUIDE TO ELECTRONIC PAYMENT SYSTEMS COSTS 3 (1998).

Even if price coherence and Gresham's law prevented the payment card industry from achieving the first-best outcome, there is no basis for devising a government intervention that has any promise of achieving a second-best outcome.

B. Single firm vs. joint venture

Nevertheless, one might be concerned that collective action by many banks could give these banks additional market power that they would not enjoy individually *and* a joint venture of banks would have the interest and ability to exercise that additional market power in setting interchange fees. For example, Frankel argues that "[t]he credit card associations (Visa and MasterCard) therefore are able to do what check and bank note clearinghouses before them could not do—enact a schedule of interchange fees governing all interbank transactions, whether the banks are located across the country or across the street from each other."[53] We first consider empirical evidence on whether merchant discount fees are higher as a result of collective action by banks. We then argue that mandating zero interchange fees could arbitrarily bias incentives toward having closed rather than open systems.

1. MERCHANT DISCOUNTS OF SMALL VS. LARGE SYSTEMS

Industry estimates suggest that merchant discounts in 1998 were about 1.6% to 1.7% for Discover, 1.8% for Visa and MasterCard, and 2.7% for American Express, and 3.0% for Diners CIub.[54] Visa's cardholder base is 1.4 times that of MasterCard, 4.4 times that of Discover, 6.6 times that of American Express, and 63 times that of Diners Club.[55] If the exercise of collective market power through the interchange fee were of serious concern, we would expect to observe that the merchant discount on Visa transactions would be significantly higher than those of its smaller competitors.[56] However, it is not.

---

53 Frankel, *supra* note 9, at 340.

54 Figures provided by: AMERICAN EXPRESS CO., *supra* note 30; Visa U.S.A.; and Telephone Interview with Diners Club personnel.

55 Relative cardholder bases are calculated based on the number of active accounts. THE NILSON REP., No. 684 (January 1999), No. 689 (April 1999).

One consequence of American Express' higher merchant discount has been a smaller merchant base, about 90% of Visa's merchant base is on a transaction volume-weighted basis. If American Express were to charge a merchant discount close to Visa or MasterCard's average discount, it would almost certainly have at least the same acceptance.

56 Some critics might argue that setting a fair price is not a defense against price fixing. The interchange fee, as we discussed in the previous section, has many procompetitive functions. Given that, we would argue that substantial evidence must exist that interchange fees pose any serious anticompetitive potential before we would consider mandating that they be set at zero.

Additional empirical evidence on interchange fees is available by considering changes over time in Visa's interchange fee. Visa has grown tremendously in size over the last three decades. If there were a significant potential that payment systems might use interchange fees anticompetitively, then we would expect that Visa might have raised interchange fees substantially. Visa's interchange fee has, however, generally decreased over this time period. Visa's interchange fee started at 1.95% in 1970 and remained unchanged until 1978 when it dropped to 1.40%.[57] It has fluctuated since then and was at 1.45% in 1998. For comparison, Visa's transaction volume in 1998 was over 150 times its transaction volume in 1970. The number of Visa issuers in 1998 was over 30 times the number of Visa issuers in 1970. By any measure, if the exercise of collective market power through the interchange fee were a serious problem, it should be of greater concern now than in 1970, yet Visa's interchange fee has decreased by over a quarter during that time.[58]

2. INCENTIVES FOR OPEN VS. CLOSED SYSTEMS

The second problem with the distinction between closed and open systems is that it would unnecessarily bias incentives toward closed rather than open systems.[59] Suppose the antitrust authorities or the courts decided it would be better to mandate a zero interchange fee for card associations rather than to run the risk of their using interchange fees anticompetitively. It seems improbable that a similar restriction would be imposed on American Express, Discover, or Diners Club. Indeed, it would be difficult to do so since these closed systems do not have interchange fees (although a regulatory cap on their merchant discounts could be imposed). We would then be in a situation where American Express could charge a merchant discount of 2.7% while the average merchant discount for Visa and MasterCard might be only about 0.5%. This might help expand Visa and MasterCard's merchant base, and American Express might need to lower its merchant discount in response, but ultimately Visa and MasterCard are disadvantaged as systems because they are constrained in the costs they can recover from merchants.

---

57 Data from Visa U.S.A.

58 Of course, changes in demand and cost conditions have also likely affected the interchange fee over time. The effects from those changes have not been factored into this comparison.

59 Howard H. Chang et al., *Some Economic Principles for Guiding Antitrust Policy Towards Joint Ventures*, 1998 COLOM. BUS. L. REV. 223; Carlton & Frankel, *supra* note 8, at 643-68.

An alternative way to view this is that Visa issuers would receive zero interchange while American Express, as an issuer, receives an implicit interchange of over 2%.[60] Opponents of interchange fees have presented no evidence that an interchange fee of zero is more efficient than the interchange fees set by card associations. They have pointed out that the checking system continued to function and grow with the par clearance system imposed by the Federal Reserve. But this does not tell us how the checking system would have developed with nonpar exchange fees. We do not know, for example, whether electronic check authorization or check truncation might have developed more quickly than they did. With zero interchange, Visa issuers could certainly try to recover their costs on the cardholder side, but there is no evidence that this would benefit consumers. Moreover, the open associations would be at a serious competitive disadvantage versus American Express with its implicit interchange of over 2%. Closed systems thus have a pricing freedom that open systems do not have. This creates a substantial, and we would argue arbitrary, incentive for payment card systems to be closed rather than open. A potential consequence of a rule against interchange fees for the open systems is to create a world in which only a small handful of for-profit closed systems can operate profitably. Open systems have provided substantial benefits to cardholders and merchants, including making it possible for smaller issuers to offer payment cards when they might not otherwise be able to.

The evidence we have reviewed suggests there is no significant likelihood that interchange fee determination by payment card associations has been used anticompetitively. Thus, there is no reason to force the open systems to set a zero interchange fee.

C. Historical evidence

Frankel compares interchange for payment cards to exchange fees that were charged in the past by banks on bank notes and checks. He argues that the historical experience from exchange fees is instructive as to the costs from permitting payment card systems to set interchange fees.

1. BANK NOTES AND COUNTRY BANKS

One of the important financial instruments in the 18th and 19th centuries was the bank note. A consumer could purchase a bank note from Bank *A* in exchange for specie (e.g., gold). The consumer could then use the bank note to purchase goods from a merchant that was willing to accept the bank note for payment. The merchant then had a number of options. First, it could use

60 If American Express has similar costs on the acquiring side as Visa acquirers (about 0.5% of transaction volume), American Express would implicitly be receiving an interchange fee of 2.7% minus 0.5%, or 2.2%.

the note to pay for its obligations with its suppliers. Second, it could present the note in person at Bank *A*, which was then obligated to pay the face amount in specie. This right to collect at par in person applies to both bank notes and to checks, which we discuss below. And third, it could deposit the note with Bank *B*. What could Bank *B* do with Bank *A*'s note? Like the merchant, Bank *B* could also present the note at Bank *A*'s counter, for which it would receive the face amount. Alternatively, Bank *B* could ship the note back to Bank *A*, which would then ship specie back to Bank *B*. Depending on the agreement reached by the two banks, Bank *A* might deduct an "exchange fee" from the face amount of the note.

Naturally, Bank *A* would prefer, all else equal, to have as large an exchange fee as possible. If Bank *B* is located next door to Bank *A*, however, it is unlikely that Bank *B* would be willing to pay anything since it could come to Bank *A*'s counter to redeem the note at par. If Bank *B* is far away from Bank *A*, then Bank *A* might be able to charge an exchange fee. There were, however, a number of constraints on the size of the exchange fee. First, if Bank *A* had local competitors, they might agree to act as agents for Bank *B* and present the note over the counter at Bank *A*. Alternatively, Bank *B* always had the option of sending its own agent to Bank *A* for redemption. In any event, some banks did charge exchange fees. Frankel argues that it was primarily country banks, especially those in one-bank towns that were able to charge exchange fees to other banks.

Frankel argues that merchants, following price coherence, would not typically charge customers different prices. That is, a customer paying with a "foreign" bank note for which there was an exchange fee would be charged the same price as a customer paying with a local bank note without an exchange fee (or a customer paying with specie). According to Frankel, country banks were able to use the phenomenon of price coherence to their advantage:

> If a local bank did possess market power, it would benefit by the use of exchange charges to the extent that distant merchants did not pass along additional fees to the bank's customers when spending the notes, but instead incorporated the charges into their overall price structures. With users of par and nonpar currencies being charged identical prices, customers using par notes paid for the exchange fees to the same extent as users of nonpar notes .... The monopolist small town bank suffered less of a decrease in demand for its notes as it raised the exchange charge than would have occurred had its own customers borne the entire incidence of its market power.[61]

61 Frankel, *supra* note 9, at 324-25.

This is one of the historical examples presented by Frankel to argue that interchange fees charged by payment card systems might be used anticompetitively. It is an interesting story. It does not, however, tell us anything about interchange fees for two reasons: (1) the story is factually questionable, and (2) it is not relevant to interchange fees.

The available historical information on merchant acceptance of bank notes suggests that nonpar notes were often not accepted at face value. The prices of notes issued by "foreign" banks were determined in a secondary market, and they reflected the risk of the issuing bank's asset portfolio, the leverage of the bank, and the time it took to carry the note back to the issuing bank for redemption.[62] In general, notes were redeemed by brokers.[63] Newspapers reporting the prices of bank notes, called bank note reporters, were published in all major cities and were also consulted in rural areas.[64] Bank note reporters were exhaustive in their coverage—they reported a price for all existing private monies in North America—and were often used by merchants to charge a discount.[65] Therefore, the hypothesis that country banks exploited price coherence and exercised geographically-based monopoly power by setting redemption charges above the cost of shipping specie, although theoretically plausible, cannot be established based on the historical evidence.

The exchange fees charged by country banks are not, moreover, relevant to the question of whether payment card systems should be permitted to set non-zero interchange fees. Let us suppose that the country banks were able to charge exchange fees on bank notes in excess of redemption costs. They would have been able to do this because par presentment over the counter was less of a constraint on their exchange fees than for a city bank where there are many other banks that could act as agents for distant banks attempting to redeem notes. Frankel does not suggest that the country banks should have been prevented from setting exchange charges, noting that "the country banks' monopsony power resulted from the superior locations they occupied, and unilateral market power exercised as a result of a firm's superior product or location is gen-

62 *See* Gary Gorton, Pricing Free Bank Notes (1998) (mimeo, University of Pennsylvania, Wharton School); and Gary Gorton, *Reputation Formation in Early Bank Note Markets*, 104 J. POL. ECON. 346 (1996).

63 *See*, *e.g.*, THOMAS S. BERRY, WESTERN PRICES BEFORE 1861: A STUDY OF THE CINCINNATI MARKET 389-90, 394, 422, 442-43, 458, 483 (1943). *See also* Gary Gorton, *Reputation Formation in Early Bank Note Markets*, 104 J. POL. ECON. 346, 355 (1996).

64 *See* Gorton, *supra* note 62, at 354 (1996).

65 *Id*. at 355, states that "Note prices in the secondary market were reported by the bank note reporters, which were consulted when unfamiliar notes were used in a transaction (. . .)." Professor Jane Knodell (University of Vermont) has told us she found that sometimes merchants specifically advertised that they would accept bank notes issued by a specific bank at a specific discount. *See* letter from Jane Knodell to authors (November 11, 1998). *See also* BERRY, *supra* note 63, at 404, who summarizes his investigation into the standard of payments employed in Cincinnati in the early 19th century in the following terms: "(...) it is apparent that Gresham's law did not apply to the situation because inferior types of money could not be passed at face value (...)."

erally protected from antitrust challenge."[66] But the unilateral market power exercised by a country bank derived simply from its freedom from par presentment over the counter by virtue of its isolated location. It suggests only that country banks have an incentive to set exchange fees in excess of the legally mandated level of zero (for over-the-counter presentment).[67] This fact provides no guidance for determining whether collective setting of exchange fees by banks might be used to exploit collective market power.

Consider the following analogy. Suppose Visa were prevented by law (like the city banks) from setting a non-zero interchange fee.[68] But suppose that American Express is permitted to charge a merchant discount (like the country banks). American Express would, of course, go ahead and charge a merchant discount. Observing this behavior would not provide any evidence as to whether Visa would or could use an interchange fee anticompetitively, or that the merchant discount charged by Visa banks would be significantly higher than that charged by American Express. Nor does this observation provide evidence as to whether mandating a zero interchange fee is socially beneficial. It tells us only that a proprietary card system (or individual country bank) has an incentive to charge more than the legal constraints imposed on other entities but not whether the legally mandated level is optimal.

2. CHECKS AND CLEARINGHOUSES

Checks, like bank notes, must be redeemed at par over the counter. Also, like bank notes, banks have in the past imposed exchange fees for redemption of checks by other, geographically distant, banks. But there is an additional twist. Associations of banks, known as clearinghouses, were formed for the purpose of clearing checks and other coordinated activity. A clearinghouse was typically composed of all banks within a city. Among other things, this gave the banks in that city the ability to attempt to set exchange fees for redemption of their checks from banks in other cities. As long as all banks located in the city belonged to the clearinghouse, there were no banks available to serve as agents for banks in other cities. According to Frankel,

> One common function of clearinghouses was to enact for all banks in a city uniform exchange charges to be assessed on all items remitted to out-of-town banks. Thus, competing city banks were able to achieve through express collusion the same market power that iso-

66 Frankel, *supra* note 9, at 329.

67 These country banks also needed to recover their shipping and other operating costs related to redemption.

68 Under competition among acquirers, this would sharply limit the merchant discount charged by Visa banks.

lated country banks had long exercised unilaterally when clearing bank notes and checks.[69]

As we discussed with bank notes, this phenomenon represents an attempt by banks to circumvent the constraints imposed on exchange fees by the legal right to obtain par clearance over the counter. The fact that clearinghouses set exchange fees for out-of-town redemption above par says nothing more than that nonpar redemption is the likely outcome of permitting clearinghouses in a world where par clearance over the counter is not mandatory.[70]

Suppose we were in a world without mandatory over-the-counter par redemption. We might likely conclude that clearinghouses would set non-zero exchange fees for checks in that world. But we do not know whether, for example, clearinghouses set higher exchange fees than banks would acting individually. Again, the issue is whether collective setting of exchange (or interchange) fees poses any anticompetitive danger, not whether banks have an incentive to evade legally mandated par clearance.

3. PAR CLEARANCE

The fact that banks, both individually and collectively, have attempted to circumvent the constraints imposed by legally mandated par clearance suggests that par clearance over the counter is not likely to be common absent legal or governmental intervention. Without legally mandated par collection over the counter, it is unclear whether banks would have moved to par exchange of bank notes.

Checks are the most prominent form of payment with par clearance. Checks have evolved to a system of near universal par clearance in the last century in the United States.[71] But this "equilibrium" came about not through market forces but as a result of government intervention, and it did not come easily. The Federal Reserve System was established by the Federal Reserve Act of 1913, with the elimination of exchange charges as one of its primary objectives.[72] The Federal Reserve's efforts to impose par clearance were hampered

69 Frankel, *supra* note 9, at 333.

70 Of course, there are also substantial efficiencies realized through economies of scale and geographic coverage afforded by clearinghouses.

71 Frankel, *supra* note 9, at 335.

72 *See* WALTER E. SPAHR, CLEARING AND COLLECTION, especially chs. VI, VII (1926); MELVIN C. MILLER, THE PAR CHECK COLLECTION AND ABSORPTION OF EXCHANGE CONTROVERSIES, especially chs. III, IV (1949); and PAUL F. JESSUP, THE THEORY AND PRACTICE OF NONPAR BANKING, especially ch. II (1967).

because banks could choose whether to join the system or not. After an initial "voluntary" phase ended unsucessfully,[73] members that chose to join the system were required to pay all checks submitted by the Federal Reserve Banks

at par. Remittance of such checks by the Federal Reserve Bank of their own district through the mail was interpreted as presentation at their own counters. Members also paid a small charge to cover the shipping and other operating costs.

Many banks did not join the Federal Reserve System at first. After all, by joining they would give up the opportunity to earn revenues from exchange fees, which were a substantial portion of total revenues for some banks.[74] To counter this, the Federal Reserve offered several advantages for members over nonmembers. Members would be participating in what was effectively a national clearinghouse for checks, with all of the cost advantages deriving from economies of scale and geographic coverage. This became an even better deal in 1918 when the Federal Reserve started to subsidize most of the costs of operating the system, eliminating most member charges.[75] In addition, the Federal Reserve relieved members of some liability for a variety of risks associated with check collection.[76] And lastly, the Federal Reserve, given the right of par presentment over the counter, used a form of note-dueling to attempt to persuade nonpar banks to join. The Federal Reserve would collect checks from nonpar banks and use agents to collect at par in person. Given these incentives, most banks joined the Federal Reserve System and became par banks, although as late as 1964 about 10% of banks were still nonpar.[77]

It is clear that the par clearance of checks came about through considerable government intervention and effort. It is also unclear as a matter of theory whether nonpar banking would have existed absent government intervention. It seems quite possible that a nonpar bank could thrive. After all, a nonpar bank would have a potential revenue source (exchange fees) that par banks would not have. That revenue could allow it to provide better service to customers by, for example, offering more attentive service, keeping longer hours or having more branches. A nonpar bank could even provide limited par redemption of its bank notes for its own customers that held deposits. For checks, a nonpar bank could offer to cash its customers' own checks at par but still maintain exchange fees on other transactions.

---

73 Under the voluntary system, members were not obligated to remit checks to Federal Reserve Banks at par.

74 JESSUP, *supra* note 72, at 47-56.

75 Hal S. Scott, *The Risk Fixers*, 91 HARV. L. REV. 737, 753 (1978).

76 *Id*. at 755-61.

77 JESSUP, *supra* note 72, at 23.

Frankel has cited Selgin and White[78] for support that par exchange of bank notes would result from competitive forces. The discussion in Selgin and White, however, does not say anything about par collection *over the counter*. Selgin and White argue that banks may reach agreement of par exchange of each others' notes (not of redemption in specie) given the existence of the right to collect at par over the counter. They point out that banks may engage in tactics such as "note-dueling" where banks collect notes issued by other banks and transport them to the issuing bank for collection at par over the counter. Such mass redemption of notes could force an unprepared issuer to suspend payments thus damaging its reputation and possibly providing a competitive advantage to the bank engaging in this tactic. Given the possibility of an equilibrium where all banks engage in note-dueling, Selgin and White argue that banks may have common incentives to avoid this by agreeing to par exchange of each others' notes. But this discussion presupposes the existence of the right to collect at par over the counter.

The common law requirement that bank notes and checks are redeemable at par over the counter is also more complicated than it seems. It is simple to say that it means that bank notes and checks clear at par, but there are a number of additional conditions that are also imposed on clearance. For example, the law also requires the merchant (or its bank) seeking to collect on a check to assume the risk that the signature was forged or that there are insufficient funds to cover the check. To mandate par clearance without specifying all of the associated conditions and obligations of the various parties would be meaningless. Therefore, if any attempt were made to prohibit payment card systems from setting non-zero interchange, it would also be necessary to spell out all the rules associated with a transaction. For example, issuers currently assume the risk of fraud and nonpayment on properly authorized payment card transactions. Instead of using interchange to apportion costs, a payment card system could shift the assumption of risk to acquirers rather than issuers. It is likely more efficient for issuers to bear the risk because they have more information on cardholders. This includes the ability to see all of a cardholder's transactions to detect patterns common to fraudulent use whereas an acquirer will generally have access to only a small fraction of a cardholder's transactions. Therefore, although there is a tendency to simply suggest eliminating interchange fees for payment cards, it must be recognized that this elimination would potentially alter the rights and obligations of all parties in the system.

78 George A. Selgin & Lawrence H. White, *The Evolution of a Free Banking System*, 25 ECON. INQUIRY 439, 446-47 (1987).

V. SUMMARY

This article has explained the economic purpose of interchange fees in payment card associations. The interchange fee equilibrates the issuing and acquiring sides of payment card systems. Payment card associations can use interchange fees to help stimulate demand on either the cardholder side or the merchant side of the system, as needed. This article has also considered arguments for prohibiting associations from setting interchange fees, and thereby requiring individual negotiations between issuers and acquirers, requiring that the associations have an interchange fee of zero. We have shown that these government interventions would likely decrease social welfare.

THE PROBLEM OF INTERCHANGE FEE ANALYSIS: CASE WITHOUT A CAUSE?

Christian Ahlborn,
Linklaters & Alliance

Howard H. Chang,
NERA Economic Consulting

David S. Evans,
NERA Economic Consulting

Previously published in European Competition Law Review, Vol. 22, Issue 8, Aug. 2001.

Interchange fees can serve the vital function of internalizing the externalities of a two-sided market. Any payment card system, be it a three-party proprietary system or a four-party bank cooperative, confronts similar factors in determining fees paid by merchants and cardholders. Any system will want to set relative prices to the two sides to ensure participation by both. Without the flexibility to adjust prices for the two sides, the success or viability of a two-sided product can be greatly reduced.

Abstract

Much of the misunderstanding of the role of interchange fees follows from the misconception that payment card services fit into a standard vertical market structure in which "upstream" issuers supply inputs to "midstream" acquiring banks, which in turn provide services to "downstream" merchants. In fact, cardholders are consumers of payments services, too, and the interchange fee accounts for the relative importance of merchants and cardholders in developing the system.

Nothing has changed in recent years to justify a radical change in the interchange fee mechanism. Individual banks still lack incentives to take into account costs and benefits that are external to them. Thus, a zero-interchange-fee rule would leave the card associations without an instrument to balance the two sides of the market. Indeed, it would greatly favor three-party proprietary systems.

In a recent article in *European Competition Law Review*, David Balto analysed the issue of interchange fees in payment card systems.[1] These are fees that banks pay one another for each credit card and debit card transaction made by their customers.[2] They arise when two banks—the merchant's bank and the cardholder's bank—are involved in the transaction.

Balto regards interchange fees as "an effective tax on merchants and ultimately consumers, that often seems unresponsive to either competition or other economic forces".[3] Various competition authorities and regulatory agencies also have concerns about interchanging fees and are currently investigating the issue.[4]

Balto's competition analysis of interchange fees, in a nutshell, runs along the following lines:

- Interchange fees may have been justified in the past on the basis that they compensated card issuing banks for certain costs that might not otherwise be recovered;
- This cost justification provided for a "narrow and tenuous exception to the traditional antitrust skepticism towards collective price fixing";
- Due to a change in the underlying technological and economic circumstances, the cost argument for interchange fees is now lacking in many respects but interchange fees have not decreased accordingly;
- Interchange fees are now unnecessarily high translating into unnecessarily high payment card costs to merchants which in turn are passed on to consumers in the form of higher retail prices; and
- It is impracticable for merchants to charge different prices for cash and card purchases, so cash users are actually subsidizing card users.

---

* *Christian Ahlborn is a competition lawyer with Linklaters & Alliance. David Evans and Howard Chang are economists with National Economic Research Associates, Inc.*

1 David A. Balto, "The Problem of Interchange Fees: Costs without Benefits?", [2000] E.C.L.R. 215.

2 ATM systems also set interchange fees. This article focuses on credit and debit card interchange fees because many of Balto's arguments are specific to credit and debit card transactions. For example, his argument that cash customers at merchants are subsidizing credit card customers as a result of credit card interchange fees would not apply to ATM interchange fees.

3 Ibid. at 222.

4 See European Commission, Commission Plans to Clear Certain Visa Provisions, Challenge Others, Press Release, October 16, 2000, available at http://europa.eu.int/rapid/start/cgi/guesten.ksh?p_action.gettxt=gt&doc=IP/00/1164|0|AGED&lg=EN; Reserve Bank of Australia and Australian Competition and Consumer Commission, *Debit and Credit Card Schemes in Australia* (October 2000), available at http://www.accc.gov.au/docs/Banks_Interchange2.pdf; Don Cruickshank (Chairman, Banking Review), *Competition in UK Banking: A Report to the Chancellor of the Exchequer* (March 2000) at 247–272, available at http://www.rroom.co.uk/response/Mon/annexd3.pdf. Although the U.S. authorities have not launched an investigation to our knowledge, it should be noted that Mr Balto is a senior official at the Federal Trade Commission.

Balto's analysis of interchange fees raises two fundamental issues. The first concerns the underlying economic rationale for interchange fees: are interchange fees really no more than a cost-compensation mechanism between different banks? The second relates to the nature of competition in the payment card sector and the impact that interchange fees have on competition: do interchange fees restrict competition, and more specifically, are they appropriately characterized as collective price fixing"?

Section I provides an overview of the payment card industry and highlights the features, such as open payment card system, network effects and two-sided products, that are critical for the understanding of interchange fees. Section II discusses the rationale for interchange fees while section III analyses the impact of interchange fees on competition. Section IV briefly deals with the alleged subsidies from cash users, an issue that raises a market-failure concern that banking regulators might consider, but does not appear to be something that competition policy regulators would ordinarily deal with.

I. PAYMENT CARD SYSTEMS: AN OVERVIEW

The payment card business started in the United States in 1950 when Diners Club introduced a card that people could use to pay for meals at associated restaurants in Manhattan.[5] Diners Club's success persuaded American Express, then a thriving travel agency and travellers cheque firm, to launch its own card brand in 1958.[6] The American Express card became the premier charge card used by business travellers by the early 1960s. Later that decade, the bank associations, Visa and MasterCard, introduced national credit cards and expanded well beyond the traditional travel and entertainment sectors. By the early 1970s, cards were a global business.[7]

The payment card business grew rapidly as customers and retailers became aware of the greater convenience: customers did not have to carry around large sums of cash or thick cheque books and could defer payment for a few weeks, while retailers faced increased demand without having to offer their own credit programmes.

5 "Dining on the Cuff", *Newsweek* 73, January 29, 1951; "Credit Card 'Pays' Entertainment Bills", *Business Week* 34, November 11, 1950. This is the origin of the general-purpose payment card industry in which cards could be used at many independent merchants. Earlier in the century, certain retailers offered cards that could be used at their stores.

6 Peter Z. Grossman, *American Express: the Unofficial History of the People Who Built the Green Financial Empire* (1987), pp. 254, 264–285.

7 For a detailed account and economic analysis of the payment card industry, see David S. Evans and Richard L. Schmalensee, *Paying with Plastic* (1999).

Proprietary systems versus open systems

Payment card systems fall into two groups: proprietary systems and open systems. Of the five major systems in the United States, Diners Club, American Express and Discover are proprietary systems while Visa and MasterCard are open systems.

A proprietary system consists of a single for-profit firm that signs up and services both cardholders and merchants, establishes the prices to charge them, operates the physical system that authorises transactions, bills cardholders and merchants, and retains the profits resulting from these activities. Proprietary systems are sometimes referred to as three-party systems.

Open systems, sometimes referred to as four-party systems, are run as co-operatives.[8] Members (which are financial institutions) vote for a board of directors, which in turn appoints the management of the co-operative. The management of the co-operative and its members play distinct roles within the open payment card system:

- The co-operative (the Visa or MasterCard organisation) is responsible for managing aspects of the card system from which all members can benefit and which no member could do on its own. This includes managing the brand (including advertising, brand positioning and brand innovation) and providing a system for authorisation and settlement of transactions involving more than one bank. The co-operative also provides for certain rules, which members have to follow. The co-operative as such does not retain profits; members' fees are set at a level at which they just roughly cover expenses (including, of course, some funds for working capital and contingencies) so that the co-operative breaks even.

- The members (for example Citibank or Chase Manhattan) are authorised to use the system's name and symbols in issuing cards and/or enrolling retailers (merchants) to accept them. Members compete with each other for services to cardholders and have total discretion in setting card fees and interest rates, as well as other parameters of their service; in the same way, members compete for services to merchants for which they set their prices (merchant discounts). Financial institutions that issue cards to consumers and provide services to cardholders are called "issuers"; financial institutions that enrol merchants and provide services to them are called "acquirers". Some institutions act as both issuer and acquirer.

8 For a description of MasterCard's predecessor's formation as a co-operative, see Gavin Spofford and Robert H. Grant, *A History of Bank Credit Cards* (1975), pp. 40–41. For a description of how Visa's predecessor became a co-operative, see Joseph Nocera, *A Piece of the Action: How the Middle Class Joined the Money Class* (1994), pp. 89–93.

Key economic characteristics of payment card systems

A payment card system provides a basic payment service for customers to pay merchants.[9] This basic payment service has two fundamental economic characteristics.

The payment card service exhibits network effects

The payment card service becomes more valuable as more people use it. Customers find a payment card service more valuable the more widely it is accepted by merchants. Merchants, in turn, find the system more valuable the more customers have (and indeed use) a card associated with a particular system.

These network effects are the *raison d'etre* for payment card systems, whose main function is to provide a uniform acceptance of their card brands: consumers know that their cards will be accepted at merchants displaying the marks for their cards; and merchants know that transactions with cards displaying a system's mark can be processed through the payment card system associated with that mark.

For open systems, a uniform acceptance of their brands (and hence the ability to benefit from the network effects) requires an "honour all cards" rule that obliges any merchant that joins a payment card system to accept for payment all of the cards that carry that system's mark. Without such a rule, the holder of a Visa card issued by Bank A would not be sure that his card could be used with a merchant which accepts Visa cards but which has been signed up by Bank B. Given that a *merchant serviced by* Bank B is required to accept *a card issued by* Bank A, Bank B then needs some assurance that it will be paid by Bank A. Thus, there must also be a requirement that Bank A will pay Bank B, on specified terms.

The payment service is a two-sided product

The payment service is a product that is only valuable if customers of each side use the product jointly. A transaction using a particular payment system takes place only when both a customer and merchant belong to and are willing to use that payment system.

9 In addition to the basic payment service, the system may provide additional services such as payment guarantees, credit facilitation and accounting, which benefit the customer, the merchant or both. For simplicity we will focus on the basic payment service.

The classic example of a two-sided product is a matchmaking service.[10] A matchmaking service has little value (to heterosexuals) if the only customers who join are men. Matchmakers try to achieve a balance of men and women. Another example of a two-sided product is Adobe's Acrobat software, which consists of a program to publish in the Acrobat PDF format and another program to read documents published in that format. People are only able to communicate using documents in Adobe format if the sender uses the Adobe publishing software and the receiver uses the Adobe reader software. Again, the value of the Adobe software can be derived only from joint use.

The particular implication of a two-sided product is that a supplier will not determine the price for each of the two elements of the product independently; rather, in setting its price for one side, the supplier will also take into account the indirect effect the price has on the other side and will maximise the overall profits for the product from both sides. So, if a matchmaking service that charges the same price to men and women finds that it has a mostly male client base, it will reduce the price it charges to women. Increasing its female client base may make the service more attractive for its male customers, which in turn may trigger a "virtuous circle" of increasing both its male and female client base, providing a service that is ultimately of higher value to all users.[11] The optimal price for a two-sided product may well involve what might loosely be characterised as "cross-subsidisation" from one side to the other. Adobe has chosen to charge for its publishing software but to give away its reader software, thus providing some assurance to purchasers of its publishing software that there will be a user base for documents in Acrobat format. Without the flexibility to adjust prices for the two sides, the success or viability of a two-sided product can be greatly reduced.

II. THE RATIONALE FOR THE INTERCHANGE FEE

The previous section has provided us with the building blocks that are needed to deal with the first of the two central issues: what is the underlying rationale for the interchange fee between acquiring banks and issuer banks? Are interchange fees, as Balto suggests, compensation paid by merchant banks for costs incurred by issuers, or are there other underlying economic forces? The answer to these questions comes in two parts.

10 Although there are many matchmaking services (*e.g.* B2B exchanges), the best known are those involving men and women. These include formal matchmaking services as well as informal ones such as singles' bars and discos.

11 Howard W. French, "Osaka Journal: Japanese Date Clubs Take the Muss Out of Mating", *New York Times*, February 13, 2001.

Optimal pricing by a proprietary payment card system

Payment services, as we have seen, are two-sided products that exhibit network effects. This means that a proprietary payment card system will set its prices to cardholders (such as card fees and interest rates) and to merchants (merchant discounts) in a way that maximises its overall profit from the system. Three factors, in particular, will influence the way it charges cardholders relative to merchants.

Elasticity of demand

Firms selling goods or services to different groups of consumers will tend to charge a higher price to the group that is less price sensitive (*i.e.* has a lower elasticity of demand). As mentioned earlier, Adobe is giving away its reader software (whose users are likely to be relatively price-sensitive) while charging for its publishing software (whose users are likely to be less price-sensitive). Another example is airlines, which charge business travellers more for their seats than leisure travellers.[12]

In the same way, the relative price-sensitivity of merchants (which is determined by the extent to which there are other payment devices and, in particular, by the extent to which they will lose sales if they do not take cards) and of consumers (which is again determined by the extent to which alternative payment devices are available) will affect a payment card system's relative pricing. The more price-sensitive consumers are relative to merchants for the demand of payment card services, the higher merchant fees will be relative to the cardholder fees.

There are two points worth highlighting here. First, any "cross-subsidisation" does not imply dominance or absence of competition. Secondly, such pricing is generally welfare enhancing; it covers the fixed costs for a good or service in a way that is least painful for cardholders and merchants overall (and which has the largest positive impact on system output).

Network effects

Unlike the usual case where a business sells to two independent groups of consumers (in our example above, business and leisure travellers), a payment card system has to take into account the interdependence of merchants and cardholders. Higher prices to merchants result in fewer merchants joining the system, which in turn makes a payment card less valuable to a payment cardholder. Higher prices to cardholders result in fewer cardholders, which in turn means that a payment card affiliation is less valuable to a merchant.

12 Michael E. Levine, "Airline Competition in Deregulated Markets: Theory, Firm Strategy, and Public Policy" (1987) 4 Yale L.J. on Reg. 393, 446-454.

The relative importance of these two network effects influences the profit-maximising price as well as the value of the payment system to society as a whole.

Costs and other factors

The costs of servicing merchants and cardholders must be taken into account.[13]

Pricing by an open payment card system

Would optimal pricing under an open card system be substantially different? In theory, the answer is no. In practice, however, open systems encounter a problem that proprietary systems do not face. Under an open system, members are free to set prices to cardholders and merchants, and the resulting merchant fees and credit card fees/interest rates are determined by competition among issuers and among acquirers. There is no reason why, as a result of these two independent competitive processes, the prices actually charged to customers and merchants should take into account the two-sided market and network effects discussed above. In fact, it is highly unlikely that the competing issuers and acquirers will take these externalities into account: they will only consider the impact of their behaviour on their profits, not the wider implication of their actions on the system as a whole.

Without a correction to this independent pricing, an open system will not be able to manage the right balance of cardholders and merchants. It would be like a matchmaking service consisting of two separate businesses in which one signed up men and the other women, with neither paying any attention to making sure there were enough men for the women and vice versa. The interchange fee provides a correction to this problem, remedying the pricing deficiency of the open system.

The function of the interchange fee

The interchange fee, often specified as a percentage of the transaction, is the amount that flows between the acquirer and the issuer for a transaction.

From the standpoint of the system, the interchange fee influences the relative prices faced by merchants and cardholders. Where, for example, the interchange fee is paid by the acquirer to the issuer,[14] the interchange fee is one of the costs that the acquirer must consider when it sets its prices to the merchant,

13 There may be other relevant factors, for example the sale of complementary products (such as credit facilitation).

14 This is the case in most—but not all—card systems.

and the acquirer will pass on some or all of this cost (depending on the nature of the competition among acquirers) to the merchant. A higher interchange fee therefore generally leads to a higher merchant discount. At the same time, the interchange fee is also one of the sources of revenue that the issuer must consider when it sets its prices to the cardholder. As a mirror image to what happens on the acquirer's side, part or all of the issuer's benefit will be passed on to the cardholder and will therefore result in lower cardholder fees.

There is another way of thinking about the interchange fee that is helpful. The total price of a card transaction is the amount of money that the card-holder and the merchant both pay. Since they jointly demand this service and the card system jointly supplies it to them, this total price really reflects what the card system is charging for the service. The specific amounts paid by the cardholder and the merchant really reflect how the system has chosen to collect this price, much as a matchmaking service collects from men versus women in the case of dating services or from buyers and sellers in the case of B2B exchanges. In open systems, the interchange fee is the mechanism that determines how that total price is divided between the two matched customers.

A common misconception about the interchange fee seems implicit in Balto's article. The interchange fee is not a price paid by the acquirers (and thus indirectly by merchants) for services rendered by the issuers. This view of the interchange fee as a price is based, erroneously, on a fictitious "vertical structure" of the industry: the "upstream issuers" supply an input to "midstream acquirers", who then supply a final service to "downstream merchants". In this vertical structure, the interchange fee is the acquirers' payment for the issuers' input and is therefore a price in that sense. But this vertical structure completely ignores the role of cardholders as consumers of the payment service; it is these cardholders that merchants get access to via their acquirers. Unlike this fictitious world, in the real world, the interchange fee affects not only the marginal cost of merchants but also the size of the cardholder clientele.

In light of the above, we can therefore conclude that interchange fees, far from being a mere compensation for certain costs as Balto suggests, are in fact a complex mechanism for ensuring the optimal functioning of an open system. The interchange fee is a device that enables the system to influence the relative merchant and cardholder prices: (i) it accounts for the relative importance of merchants and cardholders in developing the system; and (ii) it determines the extent to which cardholders and merchants will pay for the costs of the system.

Empirical evidence

The theory presented above is in accord with the facts. The relative fees charged by card systems to merchants and cardholders seem to vary consistently with the three factors identified above. The original charge card systems in the United States—American Express and Diners Club—charged merchant

discounts in the range of 5-10 per cent during their first decade. Both systems targeted cards to the travel and entertainment sector and were not initially interested in seeking widespread merchant coverage outside that sector. When the bank associations, Visa and MasterCard, entered the market and introduced national credit card products in the mid-1960s in the United States, they wanted to expand well beyond the traditional travel and entertainment sector. Not surprisingly, the merchant discounts for their products were much lower than those of American Express and Diners Club. As a result, they were able to get many more merchants to sign up for their cards. When the on-line debit systems entered the U.S. Market in the late 1980s and early 1990s, they faced a very different situation. They already had a base of cardholders that had ATM cards as part of their current accounts. Merchants, on the other hand, could not accept on-line debit without installing a new technology—pin pads. Consequently, on-line debit card systems chose a merchant discount rate that was much lower than for credit because otherwise merchants would not have installed the necessary technology.

Is there an alternative?

Individually negotiated interchange fees and "zero" interchange fees have been suggested as alternatives to the current situation. Both suggestions are fundamentally flawed.

We have shown that open payment card systems require an interchange fee (or something equivalent) because the "honour all cards" rule requires an agreement between different banks when one bank's cardholder conducts a transaction at another bank's merchant.

The interchange fees cannot be individually negotiated for two reasons: first, as mentioned above, individual members would not take into account the externalities that result from the two-sided market and network effects discussed above and hence individually negotiated interchange fees would not be effective in balancing the interests of cardholders and merchants. Secondly, individually negotiated interchange fees are not manageable from a practical point of view. In a small system with 100 member banks, 4,950 agreements would have to be negotiated. With over 21,000 member financial institutions,[15] Visa would require more than 220 million agreements. Furthermore, it is by no means clear that all members would reach an agreement. But without agreement, issuing banks could refuse to honour acquiring bank transactions and thereby "hold-up" the acquiring banks for huge interchange fees[16]; their refusal would reduce the merchant base and ultimately reduce the value of the card brand to all cardholders and merchants.

15 Visa International, *Who We Are*, available at http://www. visa.com/av/who/main.html.

16 Visa could require issuers to honour transactions but that would be tantamount to a default interchange fee of zero (or whatever rate was specified under the requirement).

The alternative proposition, namely that issuers must reimburse acquirers at par, effectively amounts to mandating an interchange fee of zero. Setting the interchange fee at this arbitrary level would remove the open systems' ability to react to cardholder/merchant imbalances and would put them at a serious disadvantage with respect to competing proprietary systems, ultimately reducing competition in the payment card industry. And, of course, setting the interchange fee at zero is just as much "collective price fixing" as setting it at any other number—so this cannot be a solution to the competition problem raised by Balto.

The only way to remove interchange fees while maintaining an efficient payment card system would be to turn open systems into proprietary systems (although merchants would still pay an "implicit" interchange fee, as can be observed from American Express which currently charges merchants higher discount than the open systems in the United States). The price of turning explicit into implicit interchange fees, however, would be high. Only a few banks could likely operate their own proprietary systems and we would be left with just a handful of issuers. Most of the current competition among members of open systems would be eliminated.

III. THE IMPACT OF INTERCHANGE FEES ON COMPETITION

This brings us to our second issue, namely the question of whether interchange fees lead to anti-competitive effects (similar to price fixing or otherwise) in the payment card industry. Two aspects have to be distinguished, namely the effects of interchange fees on intra-system competition (*i.e.* competition that takes place within the open systems) and on inter-system competition (*i.e.* competition that takes place between systems).

Intra-system competition

Unlike proprietary card systems, an open card system provides for competition among its members in most of the services rendered to cardholders and merchants. Only activities from which all members can benefit and which no member could carry out by itself are in the hands of the co-operative and are decided collectively.

The interchange fee does not provide a source of profits to the co-operative or its members. The co-operative itself does not receive the interchange fees; the fee is simply a payment from acquirers to issuers. As discussed above, the generally intense competition among issuers results in the interchange fee being mainly passed on to cardholders in the form of lower fees, while the generally intense competition among acquirers results in the interchange fee being passed on to merchants in the form of higher merchant discounts. The interchange fee does not favour a particular issuer over other issuers, or a particular

acquirer over other acquirers; it does not restrict any member's ability to compete. Furthermore, the interchange fee does not affect the intensity of competition. Ironically, one of the major complaints about the interchange fee is that it results in *too many* card transactions from a social welfare perspective.[17] Antitrust concerns typically arise in circumstances where output is too low or prices are too high. Neither circumstance is given in the context of interchange fees.

Inter-system competition

At the system level, open card systems compete with proprietary card systems (and indeed with each other). Visa and MasterCard give consumers alternatives to American Express, Diners Club and Discover, as well as other card systems that operate in particular regions (*e.g.* regional ATM systems in the United States, ecKarte in Germany, domestic debit systems in many European countries, and JCB in Japan and several other countries). They also compete with cash and cheques. Card systems compete on innovation (such as the affinity card and improvements in processing transactions), advertising and merchant acceptance.

It is at this inter-system level that interchange fees are an important competition variable. As we have seen, interchange fees allow open systems to determine the relative importance of merchants and card-holders in establishing the value of the brand, which in turn enables open systems to position themselves in the systems market in competition with each other, proprietary systems, cash, and cheques. American Express, for example, has historically sought to earn a large fraction of its revenues from merchants. It has done this by charging much higher fees to merchants in the United States than have Visa or MasterCard acquirers, and it has accepted having a much smaller number of merchants available to its cardholders as a cost of adopting this strategy. American Express has an implicit interchange fee—one that flows from the acquiring to the issuing side of its business—that is much higher than Visa's interchange fee.

Interchange fees are clearly the result of collective action by the members of an open payment card system (and they also determine the prices of merchants *relative* to cardholders—although they do not set the *absolute* prices to users of the system). In fact, by definition, any system-wide decision in an open card system is necessarily collective. For example, Visa's decision to sponsor the Olympics, MasterCard's decision to use a hologram as a security feature, and Visa's decision to invest in smart-card technology are all collective decisions.

17 See n. 2 above, at 221; Alan S. Frankel, "Monopoly and Competition in the Supply and Exchange of Money" (1998) 66 Antitrust L.J. 313, 347.

Does this mean that interchange fees (or indeed any other competitive strategy or decision by the management of a co-operative card association) are anti-competitive?

The important point to consider when answering this question is that almost no individual members could compete at the system level even in the absence of any restriction from rules of their payment card system. For inter-system competition, the strong network effects act as a significant barrier to entry for individual financial institutions and prevent a multitude of competing payment card systems. Members, linked through the "honour all cards" rule, need to manage their brand jointly. Equally, there is a collective need to balance the relative acceptance of the system by cardholders and merchants in order to promote the fullest use of the system.

If companies A, B, C and D create a joint venture to enter a market that none of them could have entered individually, then this is fundamentally different from the situation where companies E, F, G and H co-ordinate their behaviour in a market in which all of them are already present. While the latter amounts to cartelistic behaviour, the former is, if anything, pro-competitive-despite the fact that all companies engage in "collective action". In his analysis, Balto seems to confuse these two cases.

Impact of interchange fees

Therefore, interchange fees, far from being an act of "collective price fixing" are fundamentally pro-competitive. They allow an open system to compete with proprietary systems on an equal footing and to manage the system more efficiently in the view of the two-sided nature of the product and the network effects present in the market. They do not restrict output or raise total prices to cardholders and merchants.

The U.S. courts reached the same conclusion in the *NaBanco* decision in 1986.[18] None of the changes in the marketplace identified by Balto undermines that finding. Nor does the fact that, with the benefit of almost 20 additional years of economic analysis, our understanding of the role of interchange fees in two-sided markets with network effects goes beyond the classic paper by William Baxter.[19]

---

18 *National Bancard Corporation (NaBanco) v. Visa U.S.A., Inc.*, 596 F. Supp. 1231 (S.D. Fla. 1984), *aff'd* 779 F.2d 592 (1lth Cir. 1986), *cert. denied*, 479 U.S. 923 (1986).

19 See William Baxter, "Bank Interchange of Transactional Paper: Legal and Economic Perspectives" (1983) 26 J.L. & Econ. 541.

IV. CASH SUBSIDIES

Having addressed the main two issues underlying the Balto article, the remainder of the article will briefly deal with Balto's proposition that cash users are subsidising card users because it is impracticable for merchants to charge different prices for cash and card purchases. It is far from clear that merchants incur higher costs for card transactions than those using cash or cheques. But even taking Balto's assumption to be true, his argument is nevertheless flawed.

When customers use one of their cards, they impose a cost on the merchant, namely the merchant discount. Balto argues that it is hard for the merchant to charge these costs back to the customer. Card systems' association rules often prohibit surcharges on card transactions.[20] The result, Balto argues, is that customers who use cash are subsidising customers who use cards and that this results in payment cards being used too frequently. According to Balto, a zero interchange fee would be the obvious solution.

First, it is common that merchants pass along all sorts of costs that do not benefit all customers to the same extent. All customers pay higher prices when merchants offer free parking, escalators, gift wrapping, convenient store hours and many other amenities that are used by only some customers. Many merchants do not charge separately for each of these services. It is, therefore, neither surprising nor remarkable that they do not impose surcharges on credit or debit cards.

Secondly, while any of the above examples of market imperfections including the "cash subsidy" are trivial, removing the interchange fee is, as we have seen, likely to have a serious negative impact on competition: open systems would be at a competitive disadvantage with proprietary systems, such as American Express or Discover. It is questionable whether the market imperfections alleged by Balto actually exist. However, even if they did, it is not likely that curing them would be worth reducing the intense competition made possible by the existence of card associations.

Thirdly, even in the absence of reduced competition, there is no reason to believe that a zero interchange fee would improve social welfare: cardholders would pay higher prices for using their cards but would be able to use them at more merchants, which would pay lower prices for accepting cards but would have fewer customers wanting to use their cards. A mandated zero interchange fee would also prevent the associations from using interchange fees to provide incentives—for example, the associations have used incentive fees to encourage merchants to install electronic terminals.

---

20 The European Commission recently announced that it intends to take a favourable view of such rules. *See* European Commission, *Commission Plans to Clear Certain Visa Provisions, Challenge Others*, Press Release, October 16, 2000, available at <http://europa.eu.int/rapid/start/cgi/guesten.ksh?p_action.gettxt=gt&-doc=IP/00/1164|0|AGED&1g= EN>.

Finally, it should not be overlooked that it is far from clear that there is too much use of cards from a social perspective. Cash and cheques have been subsidised by the government and in some countries these subsidies continue. Moreover, in many countries consumers do not pay the direct cost of using cash and cheques and therefore tend to use them too much (in the same way Balto claims consumers use cards too much). In the United States, for example, banks do not usually charge people for taking cash out at a bank branch counter or on their ATM card on the bank's ATMs), even though the bank incurs corresponding costs. Likewise, many customers get free cheques. Card customers therefore may subsidise cash and cheque customers at the banks. So even if cash users were subsidising card users, it is far from clear that such a "subsidy" would result in excessive use of cards.

V. CONCLUSION

Suppose you were told there was a business practice that helped to create a trillion dollar industry. Suppose that this practice increased industry output. Finally, suppose that all the firms in the industry have chosen to use this practice since the beginning of the industry, regardless of whether they plausibly have market power. Such a business practice would hardly seem like a candidate for antitrust scrutiny. Yet, that is precisely what Balto has suggested.

Setting prices in order to balance cardholder and merchant demand was essential for the creation of the payment card systems, which had to deal with selling products in two-sided markets with network effects. The interchange fee has been the device used by the card associations to achieve this. It was obviously not a device for exercising market power since it is undisputed that the card associations competed intensively with cash, cheque and other payment cards in their early years in the United States. Even today, there is no dispute that in many countries, especially those in which credit cards are not as widely used, payment cards comprise a small share of transactions and compete with cash and cheques.

There is no basis for competition authorities to intervene in the setting of interchange fees. The interchange fee is not a price in the normal economic use of that term but rather a device for promoting the card brand by achieving the optimal balance of cardholders and merchants. The interchange fee determines the division of the total price of the card transaction service between the issuer and the acquirer but does not directly affect the total price. The interchange fee is set collectively, but so, too, are many matters that co-operatives must agree on to have a viable product.

There is also no basis for regulatory authorities to mandate a zero interchange fee or an interchange fee based on cost. Regulatory intervention of this sort would make sense only if the authorities could demonstrate that the current system results in a significant market failure and that either of these regulated alternatives would improve social welfare. As noted, no significant market

failure has been identified except in the trivial sense that consumers do not pay, down to the penny, for every cost they cause in the real world. Neither alternative obviously improves social welfare: reducing the interchange fee to zero would result in higher cardholder prices, lower merchant prices, fewer cardholders, and lower merchant value. There is no economic reason why all of these complex consequences balance out to an improvement in social welfare. Indeed, Rochet and Tirole at the University of Toulouse have found that, under certain circumstances, the payment card associations have private incentives to set an interchange fee at the socially optimal level (the level that an all-knowing, benevolent social planner would set).[21] That is because the associations have an incentive to balance the opposing demands of cardholders and merchants and cannot, by their structure, use interchange fees to capture supracompetitive profits. If a regulatory authority were to substitute its judgment for the associations', it would need to consider the same factors as the associations: demand elasticities, network effects, and costs. Only by coincidence would that consideration result in a socially optimal interchange fee of zero (or equal to some measure of cost).

21 Jean-Charles Rochet and Jean Tirole, *Cooperation Among Competitors: The Economics of Credit Card Associations* (April 7, 2000) unpublished manuscript.

PAYMENT SYSTEMS AND INTERCHANGE FEES

Richard Schmalensee,
MIT Sloan School of Management

Previously published in Journal of Industrial Economics, Vol. 50, No. 2, June 2002.

Abstract

This paper presents an explicit model of imperfect banking competition, in which a bank cooperative sets the interchange fee in order to maximize a weighted sum of the profits of card-issuing and merchant-servicing banks with some market power. After the cooperative has acted, the banks set prices to consumers and merchants in order to maximize their individual profits.

In a special case (but without any extreme assumptions) in this model, collective interchange fee determination maximizes output and social welfare in order to maximize the system's private value to its owners. While this does not occur in all cases, as a general matter both the privately and socially optimal interchange fees are determined mainly by differences between the demand, cost, and competitive conditions faced by card-issuing banks and those faced by merchant-servicing banks. Banks' markups are determined by the competitive conditions they face; the optimal use of the interchange fee is mainly to increase volume to the benefit of all parties.

In general, the interchange fee that maximizes private value may be above or below the fee that maximizes total system output, and if the value-maximizing fee is above (below) the output-maximizing fee, so is the welfare-maximizing fee. No cost-based approach to regulating interchange fees is guaranteed, even in theory, to enhance social welfare. This analysis reveals no economic case for requiring the interchange fee be set to zero or for prohibiting the use of any interchange fee.

1. INTRODUCTION

In a bank credit card transaction, the bank that has issued the card to the consumer is called the issuing bank or *issuer*, and the bank that processes the transaction for the merchant is called the acquiring bank or *acquirer*. When the issuer and acquirer are different, the acquirer pays the issuer an *interchange fee*, set collectively by the banks that belong to the system. Interchange fees differ among transactions of various sorts; in recent years, interchange fees in the Visa and MasterCard systems have averaged between one and two percent of transaction value. Changes in interchange fees generally affect *merchant* discounts, the fees paid by merchants to acquiring banks for processing credit card transactions. In the U.S., where acquiring is highly competitive, changes in interchange fees lead to roughly equal changes in merchant discounts.

In the U.S., collective determination of interchange fees by competing banks was found to be legal in the 1984 *Nabanco* decision.[1] This decision rested in part on the analysis presented by William Baxter [1983]. Baxter argued that because payment system volume is determined by the actions of both issuers and acquirers, and because interchange fees merely shift costs between these two sides of the system, collective determination of the interchange fee is not ordinary anti-competitive price-fixing. He showed that under perfect competition among issuers and among acquirers, the socially optimal interchange fee is generally non-zero.

Collective determination of interchange fees has recently come under renewed attack, particularly in Australia and the European Union.[2] One important element of this attack is the charge that because interchange fees are set to maximize profits of payment system members, rather than social welfare, it is appropriate to treat collective determination of interchange fees

* I am indebted to National Economic Research Associates and Visa U.S.A. for financial support. Howard Chang, David Evans, James Hunter, Jean Tirole, Kyle Bagwell, Robert Porter, anonymous referees, and, especially, Bernard Reddy provided valuable comments on earlier versions of this paper. I owe a special debt to the late William F. Baxter, who attracted me to this problem. I retain sole responsibility for imperfections and opinions.

+ Author's affiliation: Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142. USA. *email: rschmal@mit.edu*

1 National Bancard Corp. v. Visa U.S.A., Inc., 596 F. Supp. 1231 (S.D. Fla. 1984), *aff'd*. 779 F.2d 592 (11th Cir. 1986), *cert. denied*, 479 U.S. 923 (1986). For alternative views of the economics of this case, see Carlton and Frankel [1995a, 1995b], Evans and Schmalensee [1995], Frankel [1998], and Balto [2000].

2 This practice has recently been criticized by the Reserve Bank of Australia and the Australian Competition and Consumer Commission [2000] and has been formally challenged by the Competition Directorate-General or the European Commission [2000]. (See also Hehir [2000].)

as cartel behavior.[3] Some have argued that collective determination should simply be banned, though it is not obvious whether bilateral negotiations between issuers and acquirers would lead on average to lower or (as Small and Wright [2000] argue) higher fees. Others (e.g., Balto [2000] have argued that interchange fees should be set to zero by fiat or determined by regulators on the basis of system-related costs incurred by issuers and acquirers.

This paper analyzes the economic role played by the interchange fee in a payment system composed of profit-seeking, imperfectly competitive firms.[4] Two facts that served to motivate this work suggest that this role is quite unusual. First, ATM (Automatic Teller Machine) and debit card networks also generally set interchange fees collectively, but in some of these networks fees flow in the opposite direction: from issuers to acquirers.[5] A general analysis must thus be consistent in principle with interchange flows in either direction. Second, in the U.S., because American Express has always served as its own exclusive issuer and acquirer, it has nothing corresponding directly to the interchange fees of the Visa and MasterCard systems. Nonetheless, even though it has been smaller than both these systems in recent years, it has generally charged merchant discounts substantially *above* the Visa and MasterCard averages.[6]

---

3 See Balto [2000] and the references cited in the preceding footnote. In addition, Frankel [1998] and others have argued that by increasing the merchant discount, a positive interchange fee magnifies the distortion created because merchants are prevented (by credit card system rules and/or by transactions costs) from imposing surcharges on customers who use credit cards, even though they are more expensive to serve than customers who use cash or checks. (For a response to Frankel [1998], see Evans and Chang [2000].) Schwartz and Vincent [2000] have recently formalized this critique of merchant discounts, while in the model of Rochet and Tirole [2000], merchant surcharging can increase or decrease welfare. In a model in which credit cards serve to increase total transaction volume by enhancing liquidity, Wright [2000] finds that merchant surcharging tends to reduce welfare by reducing cardholding. These analyses each rest on different simplifying assumptions to permit tractable modeling of consumers and retailers (all neglect search and search costs, for instance, which are central to some models of retailing). All neglect the facts that cash and checks are regulated and subsidized and that their costs to merchants generally differ. In light of all the departures from first-best optimality in this context, the theory of the second-best suggests that regulating card system merchant discounts will raise welfare only by chance. There is even less reason to think that welfare would be increased by regulating only the discounts of the bank card systems (via attacks on interchange fees) and not those of the proprietary systems. I will, in any case, neglect these issues for simplicity in what follows.

4 Unless competition is at least slightly imperfect, it is hard to model choice of the interchange fee at the system level. Evans and Schmalensee [1995, pp. 899-901] show that in a perfectly competitive world with no frictions, any interchange fee is consistent with a zero-profit market equilibrium. In such a world, if merchants cannot give a discount for cash or impose a surcharge for credit purchases, a single interchange fee, determined by costs and equal to zero only by chance, is consistent with market equilibrium.

5 For debit card networks, see Faulkner & Gray [1999, pp. 22-26] and Reserve Bank of Australia and Australian Competition and Consumer Commission [2000]. When a customer pays with a check, no party pays anything like an interchange fee. But this zero-fee regime was produced by the Federal Reserve, not unregulated market forces: see, e.g., Spahr [l926], Jessup [1967], Baxter [1983], and Frankel [1998].

6 See Evans and Schmalensee [1999, chs. 6 and 8]. Discover/Novus operates as a proprietary system like American Express. It has generally charged lower merchant discounts than the bank card systems, though on average its discounts have exceeded the markups charged by bank card acquirers over the bank card systems' interchange fees.

The key assumption of the analysis here is that the value of a payment system to issuers is affected by the behavior of acquirers and vice versa. This network externality can only be addressed at the system level, and we show that the interchange fee provides a simple, though imperfect, tool for addressing it. The main economic role of the interchange fee is not to exploit the system's market power; it is rather to shift costs between issuers and acquirers and thus to shift charges between merchants and consumers to enhance the value of the payment system as a whole to its owners.[7] The sign and magnitude of the value-maximizing interchange fee depend on the system's objectives, on differences in costs and in demand elasticities of issuers and acquirers, differences in the intensity of competition on the two sides of the system and, in general, on differences in spillover effects between them.

Under imperfect competition, no matter how vigorous, one would not expect the interchange fee (or any other price) to be chosen in a socially optimal fashion. It is thus remarkable that under non-extreme assumptions, the privately optimal interchange fee *is* also socially optimal: I show below that it maximizes both total system output and a conventional Marshallian measure of social welfare. More generally, in deciding whether collective determination of the interchange fee should be treated like ordinary cartel price-fixing, the key question is whether collective fee setting, like ordinary price-fixing, is generally used to increase profit by reducing output. The answer is clear: it is not. The privately optimal fee may be above or below the socially optimal fee, and the difference does not turn on the level of market power.

In a paper complementary to this one, Rochet and Tirole [2000] assume perfect competition among credit card acquirers and imperfect competition among issuers and retailers. They explicitly model the retail sector, allowing for strategic behavior, and simplify by assuming identical retailers. This simplification enables them to derive welfare measures from fundamental cost and preference assumptions. Rochet and Tirole focus on equilibria in which all retailers accept credit cards, while an important feature of the analysis here is that retailer acceptance varies among equilibria. Thus the Rochet-Tirole setup facilitates rigorous welfare analysis, while the assumptions made here facilitate exploration of the balancing role of the interchange fee. Consistent with the results obtained here, Rochet and Tirole find that the profit-maximizing interchange fee never reduces the output of credit card services below the efficient level.

7 Particularly in the early years of the bank credit card systems, most banks functioned as both issuers and acquirers. The basic externality on which this analysis rests is still present, however, as long as the profits any particular bank earns from its issuing (acquiring) operations is affected by the actions or other banks' acquiring (issuing) operations.

II. BASIC ASSUMPTIONS

For simplicity, the exposition that follows concentrates on bank credit card systems, though the basic analysis applies more generally. Bank credit card systems are operated on a *cooperative* basis: they pass interchange fees through from acquirers to issuers, and they pay no dividends to the banks that own them.[8] In contrast, *proprietary* systems like American Express earn profits at the system level, whether they are *unitary* (and do all issuing and acquiring themselves) or *non-unitary* (and contract with others to do some issuing and/or acquiring). We explore some implications of these alternative structures at the end of Section IV.

Because the volume of transactions in any particular bank card system is determined by the interaction of consumers' decisions to use the card and merchants' decisions to accept it, actions of acquirers impose external effects on issuers and vice versa. Any particular card brand is more valuable to consumers the more merchants they expect to accept it, and accepting any particular card is more valuable to merchants the more consumers they expect to carry and use it. Finally, any given volume of transactions can in general be produced by an infinite number of combinations of household and merchant activity, and thus of acquirers' and issuers' efforts to stimulate demand.

Let Q^T be the value of transactions (each assumed for simplicity to have the same monetary value) on a bank card system. It is useful to begin with the simplest case of *bilateral monopoly*: a single issuer and a single acquirer. A convenient demand structure that illustrates the key system-level features discussed just above is the following:

$$(1) \qquad Q^T = Q^m\left(P^a\right)Q^c\left(P^i\right).$$

The quantity Q^m reflects merchants' willingness to accept cards; it is a decreasing function of the per-transaction price, P^a that is fixed by the acquirer and that corresponds to the merchant discount charged by all payment card systems. Similarly, Q^c reflects consumers' willingness to carry

8 See, generally, Evans and Schmalensee [1993; 1999]. In contrast, some ATM and debit card networks are proprietary and earn significant profits for their owners (Kim [1998], Faulkner & Gray [1999]). Thus some of these networks impose charges on their members that exceed system-level costs, while the bank card sys tems do not. I explore some consequences of this alternative, proprietary regime at the end of Section IV. In all that follows I assume system-level costs to be zero for simplicity; they are in fact small relative to interchange fees.

and use cards; it is a decreasing function of P^i, the effective per-transaction price charged by the issuer to consumers. P^i could take the form of average interest payments on outstanding account balances.[9] I refer to Q^m and Q^c as *partial demands* in what follows.[10]

One useful way to think of this demand structure is as follows: consumers' desired level of transactions volume is given by (1) as a function of P^i, with Q^m treated by consumers as exogenous. Here Q^m embodies *network effects*: the lower is P^a, all else equal, the greater is merchants' aggregate willingness to accept cards, thus the higher is Q^m; and the higher is Q^m, the greater the volume of card transactions desired by consumers. Similarly, merchants' demand for transactions is given by (1) as a function P^a, with Q^c treated by merchants as exogenous. On this side of the system, Q^c embodies network effects. In equilibrium the transactions volumes desired by both consumers and merchants equal the actual volume.

On this interpretation, one can use equation (1) to derive Marshallian partial equilibrium (consumers' plus producers' surplus) welfare functions for consumers and merchants. Treating Q^m as a constant, for instance, solve (1) for P^i and integrate under the resulting demand curve to obtain

$$\text{(2a)} \qquad U^c\left(Q^T;Q^m\right)=\int_0^{Q^T} Q^{c-1}\left(x/Q^m\right)dx,$$

where Q^{c-1} is the inverse of the Q^c partial demand function. Similarly, on the merchant/acquirer side of the system,

$$\text{(2b)} \qquad U^m\left(Q^T;Q^c\right)=\int_0^{Q^T} Q^{m-1}\left(x/Q^c\right)dx,$$

where Q^{m-1} is the inverse of the Q^m partial demand function. Then if C^a and C^i are the acquirer's and issuer's constant per-transaction costs, respectively, the corresponding Marshallian welfare measure is given by

$$\text{(3)} \qquad W=U^c\left(Q^T;Q^m\right)+U^m\left(Q^T;Q^c\right)-\left(C^a+C^i\right)Q^T.$$

9 In fact, total costs to both consumers and merchants may have fixed components— including annual fees, terminal installation costs, and transactions costs or dealing with an issuer or acquirer. With nontrivial fixed costs, expected per-transaction cost depends on frequency or use, which for consumers depends on expected merchant acceptance and for merchants depends on expected consumer use. I simplify by following the relevant literature and assuming that these sorts of fixed costs can be neglected in equilibrium because acceptance and use expectations are fulfilled.

10 This specification of network effects is, of course, somewhat restrictive. A supplemental appendix available at the *Journal*'s web site (www.stern.nyu.edu/~jindec) analyzes a model in which partial demand functions are linear, and each partial demand function exhibits network effects directly by being an increasing function of the expected partial demand on the other side of the system. This generalization complicates the analysis of the linear case considerably but does not change any fundamental conclusions.

This measure, of course, does not reflect the facts that merchants are not final consumers and that competition among merchants is likely to be imperfect. Nonetheless, given the popularity of Marshallian welfare analysis in a variety of policy settings—including, in particular, the analysis of regulated prices—this measure provides a potentially interesting benchmark here.

Continuing with the bilateral monopoly case for simplicity, if T is the per-transaction interchange fee, the acquirer's profit is

$$\Pi^a = \left[Q^m\left(P^a\right)Q^c\right]\left(P^a - C^a - T\right), \tag{4a}$$

where we adopt the convention that the interchange fee is positive when it is paid (as in actual bank credit card systems) from acquirers to issuers.[11] Similarly, because the interchange fee is simply transferred by the system from the acquirer to the issuer in cooperative payment systems, the issuer's profit is given by[12]

$$\Pi^i = \left[Q^c\left(P^i\right)Q^m\right]\left(P^i - C^i + T\right). \tag{4b}$$

Equations (4) show that if it were possible to shift system functions easily between issuers and acquirers, and thus to change C^i and C^a at will by the same absolute amounts but in opposite directions, there would be no need for a separate interchange fee. As Rochet and Tirole [2000] stress, however, some important functions—such as dealing with consumer default or merchant-based fraud—are more efficiently handled by one side of the system or the other. Accordingly, C^i and C^a are treated as fixed.

System behavior is modeled throughout as a two-stage game. In the second stage, the acquirer chooses P^a to maximize Π^a, treating T as fixed. Simultaneously, the issuer chooses P^i to maximize Π^i, treating T as fixed. When there are multiple issuers and/or acquirers, each takes T and all the others' Ps as fixed. These are textbook problems, with objective functions that are concave under standard assumptions.

---

11 I ignore throughout the additional complications that may arise if issuers and acquirers participate in several payment systems. In the U.S., for instance, commercial banks provide both checks and credit cards. Moreover, through an arrangement referred to as 'duality,' most U.S. banks issue both Visa and MasterCard cards. Some implications of duality are discussed in Evans and Schmalensee [1999, ch. 8].

12 I have explored generalizations of this framework in which the acquirer can invest a non-negative amount, F^a, in marketing to build merchant demand, and the issuer can invest a non-negative amount, F^i, to build consumer demand. Then the two objective functions become $\Pi^a = \left[Q^m\left(P^a, F^a\right)Q^c\right]\left(P^a - C^a - T\right) - F^{\text{'}}$, and $\Pi^i = \left[Q^c\left(P^i, F^i\right)Q^m\right]\left(P^i - C^i - T\right) - F$. Unfortunately, analysis of the choice of marketing outlays in this framework, even if price competition is assumed away, turns out to involve a high ratio of technical difficulty to added insight. Accordingly. I assume $F^a = F^i = 0$ in the text and confine discussion of results involving marketing spending to footnotes.

In the first stage, the interchange fee is chosen to maximize the system's *private value*,

$$V = \alpha \prod^{a} + (1-\alpha)\prod^{i}, \text{ for } 0 \le \alpha \le 1. \tag{5}$$

Because of the multiplicative demand structure assumed here, $V(T)$ is *not* generally globally concave, even for well-behaved partial demands. Except in pathological cases, however, V will be smooth, and values of T large enough in absolute value to drive either partial demand close to zero will not be optimal because total output will also be close to zero. Thus one of the solutions to the first-order condition $dV/dT=0$ will normally signal the global maximum.

If side payments were possible, it would be natural to set $\alpha = 1/2$ in (5) and assume maximization of total system profit. But side payments are typically not possible; $\alpha = 1/2$ is thus not necessarily descriptive of actual systems; and departures from this symmetric case are instructive. Because determining the interchange fee requires collective decision-making, which may be quite unwieldy and time-consuming, it is natural to model T as being set to maximize V in the game's first stage, before the individual banks' pricing decisions.

The next section examines what can be said about this model without specifying the functional forms of the partial demands and shows that the interchange fee plays a very different role from an ordinary market price. Section IV considers in depth the tractable case in which issuers' and acquirers' partial demand functions are linear. The welfare and output consequences of setting T to maximize private value are considered, as are the welfare and output implications of replacing cooperative systems with proprietary systems. Section V summarizes some implications of this analysis.

III. GENERAL DEMANDS

III(i). *Double Marginalization*

When there is market power on both issuing and acquiring sides of the system, as we generally assume here, there is always a form of double marginalization (or, somewhat more precisely, uncoordinated pricing of complements) in the second stage of the game described above.[13] The interchange fee cannot help with this problem.

13 This is a special case or the moral hazard problem analyzed by Holmstrom [1982].

To see the double marginalization problem, consider the impact on V of a change P^a at bilateral monopoly equilibrium:

$$(6) \qquad \frac{\partial V}{\partial P^a} = \alpha \frac{\partial \Pi^a}{\partial P^a} + (1-\alpha)\frac{\partial \Pi^i}{\partial P^a} = (1-\alpha)\left(P^i - C^i + T\right)Q^c \frac{\partial Q^m}{\partial P^a}.$$

The second equation holds because $\partial \Pi^a / \partial P^a = 0$ at equilibrium. The right-hand side of this equation will be negative as long as the issuer margin is positive, a condition that must hold in a sustainable equilibrium with non-zero issuer market power. Since the analysis is symmetric, it follows that *at a bilateral monopoly equilibrium, small reductions in P^a and/or P^i would increase V.*[14]

The interchange fee cannot contribute to the solution of this double marginalization problem. Because it can only shift costs from one side of the system to the other, the interchange fee can only mitigate problems caused by *differences* between the issuing and acquiring sides. The obvious ways to deal with the double marginalization problem are to build a unitary, proprietary system (like American Express or Discover in the U.S.) or to have competition in both issuing and acquiring (like Visa and MasterCard in the U.S.).[15]

III(ii). ***Bilateral Monopoly***

When $\alpha = 1/2$, the first-stage bilateral monopoly objective function can be written as

$$(7) \qquad V = \left\{Q^m\left[P^a(T)\right]Q^c\left[P^i(T)\right]\right\}\left\{P^a(T) + P^i(T) - C^a - C^i\right\}.$$

If $\mathrm{d}P^a / \mathrm{d}T = -\mathrm{d}P^i / \mathrm{d}T$, shifting unit cost from one side of the market to the other leaves the second term on the right of (7) unchanged, and maximization of the first term, Q^T, is necessary and sufficient for maximization of total profit. Since $\mathrm{d}P^a / \mathrm{d}T = -\mathrm{d}P^i / \mathrm{d}T = 1/2$ when demands are linear, it follows that *under bilateral monopoly and linear demand, the interchange fee that maximizes total*

14 When marketing is possible, as discussed in note 12 above, an exactly parallel analysis demonstrates that *at a bilateral monopoly equilibrium, small increases in F^i and/or F^i would also increase V.*

15 A simple example may help fix ideas. Suppose that $C^a = C^i = 0$, $Q^m = 1 - P^a$, and $Q^c = 1 - P^i$. Then with $\alpha = 1/2$, V is maximized by setting both $P^a = P^i = 1/3$, while, because of double marginalization, the equilibrium in the two-stage game has $T = 0$ and $P^a = P^i = 1/2$. Now suppose that there are two issuers, 1 and 2, facing demands given by $q_h = Q^m\left[(1/2) - (1/2)P_h + \gamma\left(P_j - P_h\right)\right]$, with $\gamma > 0$, for $h, j = 1, 2$. Note that when the issuers' prices are equal, total demand $(q_1 + q_2)$ is the same as in the monopoly case. Assume two acquirers face the same demand functions. (A generalization of this setup is analyzed in Section IV.) With zero cost, Bertrand equilibrium involves $P^a = P^i = 1/\left[2(1+\gamma)\right]$. If $\gamma = 1/2$ on both sides of the system, this implies $P^a = P^i = 1/3$, and total profit is maximized. Output is always higher under bilateral duopoly than under bilateral monopoly in this example, and for $\gamma < 1.618$ total profits are higher as well.

profit also maximizes total system output. (Section IV shows that Marshallian welfare, W, is also maximized in this case.)

When changes in T do affect the second term in (7), profit maximization does not imply output maximization. The difference between profit maximization and output maximization depends on exactly how the partial demand functions depart from linearity, and the profit-maximizing T may be above or below the output-maximizing fee. Similarly, the differences between these two quantities and the interchange fee that maximizes Marshallian welfare depend in general on the details of the partial demand functions.

Continuing to analyze the first-stage choice of T, substitute equations (4) into equation (5), assume bank-level profit maximization at the second stage, and differentiate totally with respect to T:

$$(8)\quad \frac{dV}{dT} = \alpha\left[\frac{\partial \Pi^a}{\partial T} + \frac{\partial \Pi^a}{\partial Q^c}\frac{dQ^c}{dT}\right] + (1-\alpha)\left[\frac{\partial \Pi^i}{\partial T} + \frac{\partial \Pi^i}{\partial Q^m}\frac{dQ^m}{dT}\right], \text{ or}$$

$$\frac{dV}{dT} = (1-2\alpha)Q^T + \alpha\left(P^a - C^a - T\right)Q^m\frac{dQ^c}{dT}$$

$$+(1-\alpha)\left(P^i - C^i + T\right)Q^c\frac{dQ^m}{dT},$$

where $\partial\Pi^a / \partial P^a = \partial\Pi^i / \partial P^i = 0$ under bilateral monopoly by second-stage profit-maximization. The first term on the right of this equation illustrates that if total profit is not maximized, one role played by a positive interchange fee is that of a tax levied by issuers on acquirers. The more weight the acquirer has in the system's objective function under bilateral monopoly, the less value attaches to the revenue transfer to the issuer that this tax accomplishes.

When $\alpha = 1/2$, so total profit is being maximized, equation (8) becomes

$$(9)\quad \left(P^a - C^a - T\right)Q^m\frac{dQ^c}{dT} + \left(P^i - C^i + T\right)Q^c\frac{dQ^m}{dT} = 0$$

Except in pathological cases, routine comparative statics analysis establishes that $dQ^c / dT > 0$ and $dQ^m / dT < 0$ under bilateral monopoly.[16] Now suppose

16 This statement is true in the alternative linear demand structure analyzed in a supplemental appendix available at the Journal's editorial web site, as shown there, as long as the spillover effect coefficients introduced there are not too large. When marketing is possible, as discussed in note 12, above, equation (9) still holds, but it is much harder in this case to sign these two derivatives. It is easy to use a revealed preference argument to show that if, say, Q^m is held constant and T is increased, so the issuer's effective unit cost is reduced, the monopoly issuer will change P^i and F^i as to increase Q^c. Similarly, if Q^c is held constant, an increase in T will decrease Q^m. But a decrease in Q^m lowers $\partial\Pi^i / \partial F^i$, thus lowering the optimal F^i and tending to lower Q^c. Similarly, an increase in Q^c makes acquirer marketing more attractive and thus tends to raise Q^m. There are, no doubt, stability conditions ensuring that, despite these feed backs, an increase in T will raise Q^c and lower Q^m in equilibrium when marketing is possible, but I have not attempted to derive those conditions.

that the two demand functions are identical. Then setting $T = (C^a + C^i)/2$ so that unit costs are equalized ensures that $P^a = P^i$ and that equation (9) is satisfied. Thus *under bilateral monopoly, when demand functions are identical, regardless of the level of (any measure of) collective market power, the necessary condition for profit maximization is satisfied when T is set to equalize issuer and acquirer unit costs*. (Section IV shows that this is sufficient when partial demands are linear.)

III(iii). *Other Market Structures*

When there are multiple issuers and acquirers, the expression for dV / dT is in general more complex than (8) because neither side of the system maximizes its total profit in the game's second stage. Moreover, there is no completely general guarantee that changes in the interchange fee will raise one partial demand and lower the other. Following Dixit [1986], however, the Appendix demonstrates that in a substantial class of Bertrand oligopoly models, increases in unit cost lower total market demand when standard stability conditions are imposed.[17] Thus the choice of the interchange fee under oligopoly or monopolistic competition generally involves a tradeoff between the partial demands of issuers and acquirers.

IV. LINEAR DEMANDS

This section considers the tractable case of linear partial demand functions. Suppose there are N firms on one side of the system, with demands given by

$$(10) \qquad q_i = Q^{oe}\left\{\frac{A}{N} + \Theta\left[\sum_{j \neq i} P_j - (N-1)P_i\right] - \frac{B}{N}P_i\right\},\ i = 1,\ldots N,$$

where Q^{oe} is expected partial demand on the other side of the system. (Superscripts are dropped in most of this paragraph and the next to avoid clutter.) The sum of the q_i will equal $Q^{oe}\left[A - B\overline{P}\right]$, where $\overline{P}$ is the average of the P_i. The larger is Θ, the more sensitive market shares are to differences in prices. A supplemental appendix available at the *Journal's* editorial Web site considers a generalization of this system, in which the expression in brackets in (10) depends directly on Q^{oe}. This generalization allows for more complex patterns of network effects in fulfilled expectations equilibrium, because Q^{oe} affects second-stage pricing, and adds considerable algebraic complexity, but it does not change the basic economics of the system.

17 It is easy to show that this is also true in the marketing competition model (with fixed prices) of Schmalensee [1976], when the stability conditions derived there are assumed.

Suppose all firms on this side of the market have unit cost C', net of interchange. (If this is the acquiring side of the system, $C' = C^a + T$, while $C' = C^i - T$ on the issuing side.) Multiplying (10) by $(P_i - C')$, differentiating, and solving for a symmetric equilibrium yields

$$(11) \qquad Q = \frac{(A - BC')(B + N(N-1)\Theta)}{2B + N(N-1)\Theta},$$

where Q is the total partial demand on this side of the system, the sum of the q_i / Q^{oe}, and

$$(12) \qquad P - C' = \frac{(A - BC')}{2B + N(N-1)\Theta} \ .$$

Now suppose that there are N^a acquirers, with linear demands as above and net unit costs equal to $(C^a + T)$, and N^i issuers, with linear demands as above and net unit costs equal to $(C^i - T)$. The results of the preceding paragraph imply that at a symmetric equilibrium,

$$(13a) \quad Q^m = \frac{(D^m - B^m T)(B^m + \beta^m)}{2B^m + \beta^m}, Q^c = \frac{(D^c + B^c T)(B^c + \beta^c)}{2B^c + \beta^c}; \text{ and}$$

$$(13b) \qquad P^a - C^a - T = \frac{D^m - B^m T}{2B^m + \beta^m}, \ P^i - C^i + T = \frac{D^c - B^c T}{2B^c + \beta^c} \text{ where}$$

$$(13c) \qquad \beta^m = N^a(N^a - 1)\Theta^m, \beta^c = N^i(N^i - 1)\Theta^c,$$

$$(13d) \qquad D^m = A^m - B^m C^a, \text{ and } D^c = A^c - B^c C^i.$$

The larger is $\beta^m(\beta^c)$ the more intense is competition among acquirers (issuers).

IV(i). *Output and Welfare*

From (13a), total system output is a quadratic in T, which is maximized at

$$(14) \qquad T^Q \equiv \frac{1}{2}\left(\frac{D^m}{B^m} - \frac{D^c}{B^c}\right) = \frac{1}{2}\left[\left(\frac{A^m}{B^m} - \frac{A^c}{B^c}\right) + (C^i - C^a)\right].$$

Note first that *if partial demand functions are linear and identical, it is output-maximizing to choose T to equalize unit costs*. More generally, all else equal, it is output-maximizing for interchange to flow to the high-cost side of the system, which would otherwise find it more difficult to stimulate total system demand. It is easy to show that the higher is $\left(A^h / B^h\right)$, for $h = m, c$, the lower the elasticity of Q^h, with respect to P^h at any given P^h. Thus equation (14) implies that *the more elastic the issuers' demand is relative to the acquirers' demand, the higher the output-maximizing interchange fee* (which is paid to the issuers).[18] The intuition is that it is output-maximizing to subsidize price cuts where they will do the most good to increase output for the system as a whole, and that is where demand is more elastic. Unless the partial demand functions are identical, using cost-based regulation to determine T will maximize system output only by chance.

To analyze Marshallian social welfare, W, defined by equation (3), it is first necessary to invert the demand system (8) and integrate to obtain the corresponding partial equilibrium surplus function. Confining attention to symmetric equilibria, at which the q_i, on each side of the system are equal and using (13a), we obtain

(15)

$$\begin{aligned} W &= \left[\frac{A^M}{B^m}Q^T - \frac{1}{2B^mQ^c}\left(Q^T\right)^2\right] + \left[\frac{A^c}{B^c}Q^T - \frac{1}{2B^cQ^m}\left(Q^T\right)^2\right] - \left(C^a + C^i\right)Q^T \\ &= \frac{Q^T}{2}\left\{\left[\frac{D^m\left(3B^m + \beta^m\right)}{B^m\left(2B^m + \beta^m\right)} + \frac{D^c\left(3B^c + \beta^c\right)}{B^c\left(2B^c + \beta^c\right)}\right] + \left[\frac{B^m + \beta^m}{2B^m + \beta^m} - \frac{B^c + \beta^c}{2B^c + \beta^c}\right]T\right\}, \\ &= \frac{Q^T}{2}\left(\delta + \lambda\,\mathrm{T}\right), \end{aligned}$$

where the final equality defines δ and λ. Note that $\delta > 0$, while λ has the sign of $\left[\left(\beta^m / B^m\right) - \left(\beta^c / B^c\right)\right]$. When there is only one issuer and one acquirer, $\lambda = 0$, so that *under bilateral monopoly maximizing system output is equivalent to maximizing Marshallian social welfare*. More generally, $\mathrm{d}W / \mathrm{d}T$ is a quadratic in T that with roots that resist simplification. When Q^T is maximized, however, $\mathrm{d}W / \mathrm{d}T$ has the sign of λ, and it follows easily that if T^W maximizes $W(T)$,

(16) $$\left[T^W - T^Q\right]\left[\left(\beta^m / B^m\right) - \left(\beta^c / B^c\right)\right] \geq 0.$$

18 The conclusion that only elasticity differences matter flows from the assumption that the functional forms of the two partial demand functions are the same. In general, differences in functional forms will also affect the impact on total demand of changes in the interchange fee.

The discussion below shows that the difference between T^W and T^Q reflects that fact that Marshallian welfare depends on profit as well as consumers' surplus.

IV(ii). ***Private Value***

Substituting equations (13) into equations (4) and (5), it is easy to show that private value, *V*, is proportional to

(17a)
$$V' = \frac{\omega}{B^m}\left(D^c + B^c T\right)\left(D^m - B^m T\right)^2 + \frac{1-\omega}{B^c}\left(D^m - B^m T\right)\left(D^c + B^c T\right)^2, \text{ where}$$

(17b)
$$\omega = \alpha \frac{2 + \left(\beta^c / B^c\right)}{2 + \alpha\left(\beta^c / B^c\right) + (1-\alpha)\left(\beta^m / B^m\right)}$$

Note that $\omega = \alpha$ under bilateral monopoly, when $\beta^m = \beta^c = 0$. More generally, the larger is $\left(\beta^c / B^c\right)$ or the smaller is $\left(\beta^m / B^m\right)$, the larger is ω. Differentiating (17a) yields the first-order necessary condition for first-stage maximization of private value:

(18)
$$\frac{dV'}{dT} = 2(1-2\omega)\left(D^m - B^m T\right)\left(D^c + B^c T\right) + \frac{\omega B^c}{B^m}\left(D^m - B^m T\right)^2 - \frac{(1-\omega)B^m}{B^c}\left(D^c + B^c T\right)^2 = 0.$$

When $\omega = 1/2$, equation (18) has a single real root, $T^V(1/2)$, that corresponds to a maximum of system value, and $T^V(1/2) = T^Q$, where T^Q is defined by equation (14). That is, *profit maximization under bilateral monopoly, or, more generally, private value maximization with $\omega = 1/2$, implies maximization of total system output. Under bilateral monopoly, Marshallian social welfare is also maximized.* As in Section III, the intuition is that increasing total output, by moving units costs toward equality and subsidizing price cuts where demand elasticity is high, increases the size of the pie for the system as a whole.[19]

---

19 Following the discussion in note 12, above, I have investigated a bilateral monopoly model in which prices are fixed, the issuer and acquirer choose fixed costs, and partial demands are given by $\ln\left(Q^m\right) = \phi^m \ln\left(F^a\right)$ and $\ln\left(Q^c\right) = \phi^c \ln\left(F^i\right)$, with ϕ^m and ϕ^c constants between zero and one. (The basic structure comes from Schmalensee [1976].) Numerical experiments suggest that in this model profit-maximizing interchange tends to flow, all else equal, to the side of the system with the smallest price-cost margin and to the side for which demand is more sensitive to fixed cost outlays (i.e., the side with the larger value of ϕ).

Moreover, equation (14) shows that *the privately optimal interchange fee when $\omega = 1/2$ depends only on differences between the two sides of the system, not on any measure of the level of market power*. If costs and partial demand functions are identical, for instance, the optimal interchange fee is zero no matter how much or how little market power the system as a whole enjoys. Alternatively, under bilateral monopoly with $B^m = B^c$, it is easy to show that the maximum level of system profit, a plausible measure of market power, varies with $\left(D^m + D^c\right)$, while the profit-maximizing interchange fee, $T^V(1/2) = T^Q$, varies with $\left(D^m - D^c\right)$.

When $\omega \neq 1/2$, equation (18) has two real roots. The root corresponding to a maximum of V is

$$(19) \qquad T^V(\omega) = T^Q + \left(\frac{D^m}{B^m} + \frac{D^c}{B^c}\right)\frac{1 - \sqrt{1 + 12\left(\omega - \frac{1}{2}\right)^2}}{12\left(\omega - \frac{1}{2}\right)}$$

Since $T^V(\omega)$ is a decreasing function, from (17b) *the private value-maximizing interchange fee is a decreasing function of α and $\left(\beta^c / B^c\right)$, and an increasing function of $\left(\beta^m / B^m\right)$*.

Under profit maximization, when $\alpha = 1/2$, $(\omega - 1/2)$ has the sign of $\left[\left(\beta^c / B^c\right) - \left(\beta^m / B^m\right)\right]$.

Comparing (16), if T^Π is the value of T that maximizes total system profit,

$$(20) \qquad \left[T^\Pi - T^Q\right]\left[T^W - T^Q\right] \geq 0$$

That is, *the profit-maximizing T departs from T^Q in the same direction as the welfare-maximizing T*. This result, which echoes the relation between Ramsey pricing and monopoly price discrimination and similarly reflects the inclusion of profits in the Marshallian welfare measure, does not seem to be easily generalized beyond the linear case. From (13b), with linear partial demands the more intense is competition on either side of the system, the less sensitive is the unit markup on that side of the system to changes in T. (In the limit as β increases, unit markup goes to zero, independent of T.) It is easy to show that the derivative of total system markup, $\left[\left(P^a + P^i\right) - \left(C^a + C^i\right)\right]$, with respect to T has the sign of $\left[\left(\beta^c / B^c\right) - \left(\beta^m / B^m\right)\right]$. Thus if $B^m = B^c$, for instance, and there is more intense competition on the issuing side than on the acquiring side $\left(\beta^m < \beta^c\right)$, it is both profit-maximizing and welfare-maximizing to reduce T below the output-maximizing level in order to increase total system markup.

When $\alpha \neq 1/2$, the interchange fee is affected by the desirability of shifting profit from one side of the system to the other. Under bilateral monopoly, with $\omega = \alpha$, the second term on the right of (19) directly reflects the use of the interchange fee to transfer profit from one side of the system to the other.

When $\alpha < 1/2$, for instance, so that the issuer's profit is weighted more heavily than the acquirer's, this second term is positive. In this case T is increased, all else equal, in order to transfer profit to the issuer, and, all else equal, system output and welfare are reduced as a consequence.

IV(iii). *Alternative System Structures*

In the U.S., banks' voting power in the Visa and MasterCard associations is more sensitive to issuing volume than to acquiring volume, indicating $\alpha < 1/2$. In addition, the acquiring side of the U.S. bank credit card business involves little or no product differentiation and is generally viewed as highly competitive, indicating β^m is large.[20] From equation (17b), this suggests that the polar case $\omega = 0$ is of particular interest.[21] In this case, equations (14) and (19) directly imply

(21)
$$T^V(0) = \frac{1}{3}\left[\left(2\frac{A^m}{B^m} - \frac{A^c}{B^c}\right) + \left(C^i - 2C^a\right)\right] > T^Q.$$

Note that $T^V(0)$ is independent of β^c and thus of the intensity of competition among issuers. Even though in this polar case acquirers' cost and demand conditions are weighted more heavily than those of issuers, differences between the two sides of the system remain central, and the qualitative impacts of changes in cost and demand conditions are essentially the same as under output maximization.

To evaluate the importance of this extreme departure from output maximization in a cooperative system, let Q^{MAX} be the maximum value of total system output:

$$Q^{MAX} = \left[\frac{\left(B^m + \beta^m\right)\left(B^c + \beta^c\right)}{\left(2B^m + \beta^m\right)\left(2B^c + \beta^c\right)}\right]\left[\frac{\left(D^m B^c + D^c \beta^m\right)^2}{4B^m B^c}\right]$$

(22a)
$$= K\left[\frac{\left(D^m B^c + D^c B^m\right)^2}{4B^m B^c}\right],$$

where the second equality defines K. In general K is between 1/4 and one, depending on competitive conditions among issuers and acquirers. When β^m is large, as was assumed in deriving (21), K is between 1/2 and one.

20 Structurally, the acquiring business does not look perfectly competitive. (See, generally, Evans and Schmalensee [1999, ch. 6].) Most U.S. banks contract out this function to specialists, and, because of scale economies in transaction processing, concentration in acquiring is relatively high. Still, competition in this commodity business is generally described as intense, and margins are small relative to, e.g., interchange fees, so that perfect competition may be a good behavioral approximation.

21 As noted above, this is in effect the case on which Rochet and Tirole [2000] focus.

Substitution of (21) into equations (13a) yields total system output when $T = T^V(0)$:

(23) $$Q^V(0) = (8/9)Q^{MAX}.$$

That is, *when the interchange fee is at the highest value consistent with private value maximization, total output is reduced by about 11 percent from its maximum value*.

To put this reduction in perspective and to shed light on some current controversies, it is useful to consider total output under alternative system structures. Consider first a non-unitary proprietary system, which charges acquirers a fee T^a per transaction, charges issuers a fee T^i per transaction, and sets these fees to maximize $(T^a + T^i)Q^T$. It is straightforward to show that the corresponding total output level is given by

(24) $$Q^{PN} = (4/9)Q^{MAX} = (1/2)Q^V(0).$$

That is, *moving from a cooperative system to a non-unitary proprietary system, keeping the numbers of (independent) issuers and acquirers constant, reduces output by between 50 and 56 percent*. This result should make clear the fundamental economic difference between an interchange fee passed from acquirers to issuers in a cooperative system and an ordinary per-transaction fee set by a proprietary system to maximize its profit.

Finally, consider a unitary proprietary system, which does all its own issuing and acquiring and sets P^a and P^i to maximize total system profit. The corresponding total output level is given by

(25) $$Q^{PI} = (1/K)Q^{PN} = (4/9K)Q^{MAX} = (1/2K)Q^V(0).$$

Except in the case of perfect competition in issuing and acquiring (when $K = 1$), total output for a unitary proprietary system exceeds that for a non-unitary proprietary system, all else equal. The non-unitary system's profit is, in effect, the receipts from taxing an imperfectly competitive market, thus giving rise to a double marginalization problem. On the other hand, *if competition in issuing and/or acquiring is vigorous, so $K > 1/2$, a unitary proprietary system always has lower total output than a cooperative system, all else equal*.

V. IMPLICATIONS

The policy question motivating this paper is whether antitrust authorities should condemn collective determination of interchange fees for the same reasons they would condemn competing banks fixing credit card interest rates or annual fees.[22] The analysis here provides no support for such a policy. The interchange fee is not an ordinary market price; it is a balancing device for increasing the value of a payment system by shifting costs between issuers and acquirers and thus shifting charges between consumers and merchants.[23] The first-order effect of fixing an ordinary price is to harm consumers by reducing output, while in a non-extreme case, collective interchange fee determination *maximizes* output and Marshallian welfare in order to maximize the system's private value to its owners.

More generally, our analysis shows that both the private value-maximizing interchange fee and the output-maximizing fee are determined mainly by *differences* between issuers and acquirers; symmetry makes a zero interchange fee optimal. The model employed here is thus consistent with collectively determined interchange flowing to either issuers or acquirers, and we observe both patterns in reality. We find that the private value-maximizing interchange fee may be above or below the output-maximizing fee and that the welfare-maximizing fee differs from the output-maximizing fee in the same direction as the profit-maximizing fee does. Increasing the interchange fee from its privately optimal level may increase total system output, and decreasing it may decrease output.

Even in the special case of linear partial demands, our analysis reveals no straightforward policy toward the interchange fee that can reliably be expected to improve system performance on balance.[24] Small and Wright [2000] have argued that moving interchange fees from collective determination to bilateral negotiations could raise fees on average, with unpredictable impacts on output and welfare. Similarly, if interchange fees were set to zero, nothing in this analysis suggests that total system output or welfare would be more likely to rise than to fall. Except in *very* special circumstances, no cost-based approach

22 As discussed in note 3, above, the formal analysis here does not deal with the argument that merchant discounts should be reduced (at least in part by putting pressure on interchange fees) in order to reduce distortions in retail pricing.

23 Balto [2000] and others who condemn collective determination of interchange fees seem to ignore this balancing role. Thus they condemn interchange fee increases because they raise merchants' costs and forget that the same logic says that interchange fee increases lower consumers' costs.

24 As discussed in note 3, above, some observers contend that because retailers partially shift merchant discounts to consumers using cash and checks, value-maximizing credit card systems may set interchange fees inefficiently high. Even if this argument were generally correct, despite the second-best issues raised in note 3 and the complexities discussed in the text, substantial reductions in interchange fees may well reduce card system output substantially, directly harming consumers. Thus it does not follow that reducing interchange fees to zero (or some cost-driven level) can be expected to make consumers better off on balance.

to regulating interchange fees can guarantee to increase system output or Marshallian welfare over private value-maximizing levels. It is highly unlikely that regulators would ever have enough information to implement the socially optimal interchange fees discussed in Section IV and the supplemental appendix available at the *Journal's* Web site, and these solutions rest on a set of restrictive assumptions.

Despite these uncertainties, any serious restriction on collective interchange fee determination would have one clear effect: it would make it harder for the bank card systems to compete effectively with American Express and other proprietary payment systems. As I noted above, because within the U.S. it does all its system's issuing and acquiring, American Express has been free to set merchant discounts and cardholder fees there without fear of antitrust attack. It has generally chosen to set merchant discounts that could be matched by the bank card systems only if they were to raise their interchange fees substantially.[25] In some other countries it has operated as a non-unitary proprietary system. Because the fees specified in its contracts with independent issuers and acquirers were not the result of agreements between competitors, however, they have also been immune to antitrust attack.

Barring collective interchange fee determination would create strong incentives for large institutions to abandon the cooperative bank card systems and create proprietary systems. As the analysis of Section IV indicated, however, all else equal a movement from cooperative to proprietary systems is likely to reduce total system output. All in all, there is no economic defense for an antitrust policy favoring proprietary payment systems over cooperative payment systems pursuing broadly similar strategies.

25 Discover/Novus has charged a lower average merchant discount than Visa or MasterCard acquirers, though its average merchant discounts have exceeded the markups charged by bank card system acquirers over the bank card systems' interchange fees and thus likely have exceeded the levels that would emerge in the bank card system if interchange fees were forced to zero.

APPENDIX

Consider a market with N firms selling differentiated products in which firm *i*'s demand, q_i, depends only on its price, p_i, and the *average* price of its *N*–1 rivals, p_{-i}. (As Dixit [1986. p. 119] notes, without some restrictive assumption of this sort, it is generally not possible to do comparative statics in differentiated product oligopolies.) If c_i is firm *i*'s constant unit cost, the set of first-order conditions that must be satisfied at a Bertrand equilibrium is

(A1) $$\mu_i = q_i + (p_i - c_i)\sigma_i = 0,\ i = 1,...,N,$$

where $\sigma_i \equiv \partial q_i / \partial p_i < 0,\ i = 1,...,N.$ Key quantities in the analysis that follows are

(A2) $$a_i \equiv \partial \mu_i / \partial p_i < 0,\ i = 1,...,N;\ \text{and}$$

(A3) $$b_i = \frac{\partial \mu_i}{\partial p_j} = \frac{\sigma_{-i} + (p_i - c_i)\dfrac{\partial \sigma_i}{\partial p_{-i}}}{N-1}, i = 1,...,N, j \neq i,$$

where $\sigma_{-i} \equiv \partial q_i / \partial p_{-i}, i = 1,...,N.$ The sign of the a_i, follows from the second-order conditions. The natural assumption that competing products are (gross) substitutes implies that $\sigma_{-i} > 0$, but there is no obvious reason why $\partial \sigma_i / \partial p_{-i} = \partial \sigma_{-i} / \partial p_i$, should be positive or negative. I assume that the σ_{-i} terms dominate, so that $b_i > 0$ for all i.

In order to do comparative statics in oligopoly models, it is generally necessary to invoke stability conditions to replace the 'off-diagonal' second order conditions that arise in monopoly models. (See Schmalensee [1976], Dixit [1986], and the references they cite.) Here, following Dixit [1986, p. 117], 1 assume the following diagonal dominance condition is satisfied:

(A4) $$a_i + (N-1)b_i < 0, i = 1,...,N.$$

This is a sufficient condition for stability under conventional dynamic assumptions.

Now suppose that c_i is replaced throughout by $(c_i + \theta)$. The goal here is to sign $\mathrm{d}Q / \mathrm{d}\theta$ at $\theta = 0$, where Q is the sum of the q_i. I do this by showing that a small increase in θ raises all prices in equilibrium. Letting 'dx' be shorthand for '$\mathrm{d}x / \mathrm{d}\theta$ at $\theta = 0$', totally differentiate the first-order conditions (A1) to obtain

(A5) $$a_i(dp_i)+[(N-1)b_i](dp_{-i})-\sigma_i=0, i=1,...,N.$$

Without loss of generality, suppose $dp_1 \le dp_2 \le ...dp_N$. This implies that $dp_{-1} \ge dp_{-2} \ge ...dp_{-N}$.

To show that all the dp_i, are positive under the above assumptions, let us suppose $dp_1 \le 0$ and show a contradiction. Since $a_1 < 0, b_1 > 0$, and $\sigma_1 < 0$ equation (A5) shows that $dp_1 \le 0$ implies $dp_{-1} < 0$. It then follows from the inequalities just above that all other dp_{-i}, must also be negative.

Dividing (A5) by a_i and summing across all firms in the market yields

(A6) $$\sum_{i=1}^{N} dp_{-i}\left[1+\frac{(N-1)b_i}{a_i}\right]=\sum_{i=1}^{N}\frac{\sigma_i}{a_i}.$$

The summation on the right is positive, and so, from (A4), are each of the terms in brackets on the left. It is accordingly not possible for (A6) to hold if all the dp_{-i} are all negative. The assumption that one or the prices does not increase has thus led to a contradiction, so all prices must rise when unit costs increase across the board, and total output must accordingly fall.

REFERENCES

Balto, D. A., 2000, 'The Problem of Interchange Fees: Costs without Benefits?', *European Competition Law Review*, 21, pp. 215-224.

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

Carlton, D. W. and Frankel, A. S., 1995a, 'The Antitrust Economics of Credit Card Networks', *Antitrust Law Journal*, 63, pp. 643-668.

Carlton, D. W. and Frankel, A. S., l995b, 'The Antitrust Economics of Credit Card Networks: Reply to Evans and Schmalensee Comment', *Antitrust Law Journal*, 63, pp. 903-915.

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

Competition Directorate-General, European Commission, 2000, 'Commission Plans to Clear Certain Visa Provisions, Challenge Others', Press Release IP/00/1164, October16.

Dixit, A. K., 1986, 'Comparative Statics for Oligopoly', *International Economic Review*, 27, pp.107-122.

Evans, D. S. and Schmalensee, R. L., 1993, *The Economics of the Payment Card Industry* (National Economic Research Associates, Cambridge, MA).

Evans. D. S. and Schmalensee, R. L., 1995, 'Economic Aspects of Payment Card Systems and Antitrust Policy Toward Joint Ventures', *Antitrust Law Journal*, 63, pp. 861-901.

Evans, D. S and Schmalensee, R. L., 1999, *Paying with Plastic* (MIT Press, Cambridge, MA).

Evans, D. S., Schmalensee, R. L. and Chang. H., 1998, 'Some Economic Principles for Guiding Antitrust Policy Towards Joint Ventures', *Columbia Business Law Review*, pp. 223-329.

Faulkner & Gray, 1999, *Debit Card Directory*, 1999 Edition (Faulkner & Gray, Inc., New York).

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

Hehir, G., 2000, 'Politics & Economy: EU Launches Investigation of Visa Practices—Regulators Say Certain Fees May Violate Antitrust Rules', *Wall Street Journal Europe*, October 17.

Holmstrom, B., 1982, 'Moral Hazard in Teams', *Bell Journal of Economics*, 13, pp. 324-340.

Jessup, P., 1967, *The Theory and Practice of Nonpar Banking* (Northwestern University Press, Evanston).

Kim, J., 1998, 'The Impact of Proprietary Positions and Equity Interest in the Pricing of Network ATM Services', chapter in unpublished dissertation, Department of Economics, Massachusetts Institute of Technology, Cambridge.

Reserve Bank of Australia and Australian Competition and Consumer Commission, 2000, 'Debit and Credit Card Schemes in Australia: A Study of Interchange Fees and Access,' mimeo, Canberra.

Rochet, J.-C. and Tirole, J., 2000, 'Cooperation Among Competitors: The Economics of Payment Card Associations', mimeo, University of Toulouse.

Schmalensee, R., 1976, 'A Model of Promotional Competition in Oligopoly', *Review of Economic Studies*, 43, pp. 493-507.

Schwartz, M. and Vincent, D. R., 2000, 'The No Surcharge Rule in Electronic Payments Markets: A Mitigation of Pricing Distortions?', mimeo, Georgetown University.

Small, J. and Wright, J., 2000, 'Decentralized Interchange Fees in Open Payment Networks: An Economic Analysis', mimeo, NECG and University of Auckland.

Spahr, W., 1926, *The Clearing and Collection of Checks* (The Bankers Publishing Co., New York).

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

JOURNAL OF INDUSTRIAL ECONOMICS

Published first in 1952, the *Journal* has a very wide international circulation, and is recognized as a leading journal in the field of industrial economics. It was founded to promote the analysis of modern industry, particularly the behaviour of firms and the functioning of markets.

Contributions are welcomed in all areas of industrial economics including: organisation or industry, applied oligopoly theory, product differentiation and technical change: theory of the firm and Internal organisation: regulation, monopoly, merger and technology policy. Necessarily these subjects will often draw on adjacent areas such as international economics, labour economics and law.

The Journal has a tradition of publishing a blend of theory and evidence. Theoretical papers are welcomed and should be presented so as to highlight their implications for policy and/or empirical analysis. Likewise, empirical papers should have a sound theoretical base; and where novel econometric

techniques are applied these should be clearly explained. Case studies should be motivated by, and inform, economic theory and should avoid pure description.

The Editors are ready to publish shorter notes which report significant new data or empirical results, or are short comments on subjects which have featured in previous issues of the Journal. Books are not reviewed, but substantial review articles will be considered for publication.

Information for Contributors

Authors who would like a paper considered for publication in the Journal should submit four copies (preferably printed two-sided) that include an abstract of not more than 100 words. If a paper is accepted, the author will be asked to prepare it in accordance with the Journal style guide. Papers submitted must not simultaneously be under consideration at another journal.

Typescripts from North America should be sent to the General Editor, Journal of Industrial Economics. Department of Economics, Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012-1126, USA.

Typescripts from all other countries should be sent to the Editor, Journal of Industrial Economics, Department of Economics, University of Essex, Wivenhoe Park, Colchester C04 3SQ, UK.

Electronic Mail Addresses:	(UK) jindec@essex.ac.uk (US) jindec@stern.nyu.edu
Editorial Web Page:	http://www.stern.nyu.edu/~jindec
Publisher's Web Page:	http://www.blackwellpublishers.co.uk

The Journal of Industrial Economics is published four times a year in March, June, September and December by Blackwell Publishers Limited, l08 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Maiden, MA 02145, USA.

Information for Subscribers

New orders and sample copy requests should be addressed to the Journals Marketing Manager at the publisher's address above (or by email to jnlsamples@blackwellpublishers.co.uk, quoting the name of the journal). Renewals, claims and all other correspondence relating to subscriptions should be addressed to Blackwell Publishers Journals, PO Box 805, 108 Cowley Road, Oxford OX4 1FH, UK (tel: +44(0)1865 244083, fax: +44(0)1865 381381 or email: jnlinfo@blackwellpublishers.co.uk). Cheques should be made payable to Blackwell Publishers Ltd.

2002 Subscription Prices:

	UK/Europe	Rest of World	The Americas
Institutions	£116.00	£126.00	$167.00
Individuals	£30.00	£34.00	$59.00
Students	£17.00	£17.00	$26.00

The reduced rates will be allowed to any individual student for a maximum period of three years.

Contributors to the Issue

RICHARD SCHMALENSEE: Sloan School of Management, MIT, USA.

NICHOLAS G. RUPP: East Carolina University, USA.

CURTIS R. TAYLOR: Duke University, USA.

MARIO FORNI: University of Modena and Regio Emilia, Italy.

SERGIO PABA: University of Modena and Regio Emilia, Italy.

MICHELLE HAYNES: University or Warwick, UK.

STEVE THOMPSON: University of Leicester, UK.

MIKE WRIGHT: University of Nottingham, UK.

JOSH LERNER: Harvard Business School, USA.

JEAN TIROLE: Institut d'Economie Industrielle, France.

US Mailing: Periodicals postage paid at Rahway, New Jersey. Postmaster: Send address corrections to: Journal of Industrial Economics, c/o Mercury Airfreight International Ltd Inc., 365 Blair Road, Avenel, NJ 07001, USA (US Mailing Agent).

Internet: For information on all Blackwell Publishers books, journals and services, log onto URL: http://www.blackwellpublishers.co.uk.

Back Issues: Single issues from the current and previous two volumes are available from Blackwell Publishers Journals. Earlier issues may be obtained from Periodicals Service Company, 11 Main Street, Germantown, NY 12526. USA. Tel:00l 518 537 4700, Fax:001 518 537 5899, Email psc@backsets.com

Microform: The journal is available on microfilm (l6mm or 35 mm) or l05 mm microfiche from the Serials Acquisitions Department, Bell & Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346. USA.

Advertising: For details please contact Andy Patterson, Wheatsheaf House, Woolpit Heath, Bury St Edmunds, 1P30 9RN, UK. Tel. 01359 242375, Fax 01359 242837.

© 2002 Blackwell Publishers Ltd (a Blackwell Publishing Company). All rights reserved. With the exception or fair dealing for the purposes or research or private study, or criticism or review, no part of this publication may be reproduced, stored or transmitted in any form or by any means without the prior permission in writing from the copyright holder. Authorization to photocopy items for internal and personal use is granted by the copyright holder for libraries and other users of the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, USA (www.copyright.com), provided the appropriate fee is paid directly to the CCC. This consent does not extend to other kinds or copying, such as copying for general distribution for advertising or promotional purposes, for creating new collective works or for resale. Institutions with a paid subscription to this journal may make photocopies for teaching purposes free of charge provided such copies are not resold.

COOPERATION AMONG COMPETITORS: SOME ECONOMICS OF PAYMENT CARD ASSOCIATIONS

Jean-Charles Rochet,
Toulouse University

Jean Tirole,
Institut d'Economie Industrielle

Previously published in Rand Journal of Economics, Vol. 33, No. 4, Winter 2002.

This paper models explicitly the behavior of all actors in a four-party payment system. This structural approach requires fairly strong simplifying assumptions for tractability but permits fully rigorous analysis of bank, consumer, and merchant behavior and of the determinants of the relation between market equilibrium and social welfare.

Abstract

It demonstrates that if merchants cannot offer cash discounts and if the alternative payment system is provided and priced efficiently, the interchange fee is either socially optimal or leads to an overprovision of credit card services. Even if interchange fees were too high, this shows how difficult it would be to attempt to improve matters in practice. In this simplified model, there is no guarantee that setting the interchange fee equal to zero or basing it on the costs of issuing and acquiring banks would produce a gain in social welfare. Socially optimal interchange fees depend on difficult-to-measure benefits to consumers and merchants, as well as on the exact nature and intensity of competition among issuing banks and among merchants.

1. INTRODUCTION

The rapid growth of payment cards[1] usage is a striking feature of modern economies. The payment activity, a fundamental dimension of the payment card industry, is characterized by the existence of strong network externalities: In a payment card transaction, the consumer's bank, called the issuer, and the merchant's bank, the acquirer, must cooperate in order to enable the transaction. Two successful not-for-profit joint ventures, Visa and MasterCard,[2] have designed a set of rules to govern the "interconnection" between their members:

(1) *Interchange fee:* The acquirer pays a collectively determined interchange fee (the analog of an access charge in telecommunications) to the issuer.[3]

(2) *Honor-all-cards rule:* Affiliated merchants must accept the card of any issuing member.

(3) *No-surcharge rule:* Affiliated merchants are not allowed to impose surcharges on customers who pay with a card.[4]

Some of these institutional features have gained wide acceptance. To see the benefits of a centrally-determined interchange fee cum the honor-all-cards rule, it suffices to envision the complexity of bilateral bargaining among thousands of banks as well as the cost for issuers (respectively, merchants) of informing consumers about the set of merchants (respectively, banks) with whom an agreement has been reached.[5] The latter transactions costs could be avoided by keeping the honor-all-cards rule while letting issuers and acquirers set their pairwise interchange fees. However an individual issuer would then be able to impose an arbitrarily high interchange fee since the acquirer would

---

1 Payment card is a generic name that includes credit cards, debit cards and charge cards. Since we focus here on the payment activity, these differences are immaterial for our purposes. See Chakravorti-Emmons (2001) and Chakravorti-To (2000) for discussions of the credit functionality.

2 Visa and MasterCard are each owned jointly by thousands of banks and handle 75 percent of the total volume of general purpose payment card transactions. There also exist "proprietary systems" such as American Express, in which the issuer and the acquirer are the same firm. We refer to Evans-Schmalensee (1999) for an excellent overview and analysis of the industry. Interestingly, MasterCard filed in August 2001 with the SEC to convert itself from a member-based association to a shareholder-owned company, MasterCard International, to differentiate itself from Visa.

3 In practice, the issuer guarantees the payment. The payment guarantee by the issuer can be motivated by considerations of delegated monitoring. The interchange fee depends on the fraud-control devices installed at the merchant's premises.

4 This is the US version of the rule. Interestingly, US merchants rarely offer discounts for cash payments, even though such discounts are not prohibited! In some European countries, payment card associations impose a stricter rule and prohibit any form of discrimination.

5 For more on transaction costs, see Evans-Schmalensee (1995, p 886, 887 and 890).

then face the grim choice between accepting this fee on a fraction of his payments and exiting the industry altogether. Individual issuers would become bottlenecks and their free riding would dissuade acquirers from entering the industry.[6]

In contrast, although they have not yet been successfully challenged in court,[7] two features of these interconnection rules have recurrently been viewed with suspicion by competition authorities and by some economists.[8] First, the no-surcharge rule is sometimes viewed as an attempt by payment card systems to leverage their market power by forcing more card transactions than is efficient. Second, the collective determination of the interchange fee is regarded by some as a potential instrument of collusion: Aren't the banks able to inflate payments to each other and, in fine, tax merchants and consumers? Shouldn't the access charge be regulated as in telecommunications? Even if one accepts the existence of an interchange fee, one may still be legitimately concerned that it be set too high. At a general level, agreements among competitors can be anticompetitive and one must investigate whether this is indeed the case in our context. For example, a joint venture among competitors whose primary motive is to raise the price on the final good market by overcharging for a common input and redistributing the proceeds among its members is anticompetitive. The prototypical example is that of a patent pool in which product market competitors pool their substitute patents and tax each other through proportional royalties paid to the pool, which are then redistributed to pool members as dividends (Priest 1977). Similarly, a high reciprocal access charge negotiated between rival telecommunications networks may in some circumstances be anticompetitive.[9] It is tempting to draw an analogy between such situations and that of a collectively-chosen interchange fee. As we will see, one should refrain from making such a quick transposition.

This paper analyzes the validity of these two concerns. In order to provide a policy analysis, it develops a normative framework of the determination of an efficient interchange fee and of the impact of the no-surcharge rule. The strength of our approach relative to the previous literature (initiated by Baxter 1983) is that we endogenize consumer and merchant behavior and are therefore able to perform a proper welfare analysis. This literature focused on the

6 Small and Wright (2000) analyze the inefficiencies that would result from a decentralized setting of interchange fees.

7 See in particular National Bancard Corp. v. Visa USA, Inc. 596 F. Supp 1231 (SD Florida 1984).

8 E.g., Frankel (1998) and Carlton-Frankel (1995).

9 See, e.g., Laffont et al (1998 a,b) for a formalization of this argument and a number of qualifications to it.

technological benefits to consumers and merchants brought about by the use of payment cards relative to alternative means of payment. Namely, it assumed that consumers and merchants adopt the card as long as the technological benefits exceed their payments to the banks. This ignored the fact that consumers and merchants are strategic players:

First, when (at least some) consumers know which stores take payment cards before they select which to patronize, or may leave the store when they discover the card is not accepted, card acceptance is used by merchants to attract customers. A merchant's total benefit, and thus his decision of whether to accept a card then depend not only on the merchant's technological benefit (fraud control, theft protection, speed of transactions, customer information collection,...), but also on the product of its increase in demand due to system membership and its retail markup.[10] Thus, the earlier literature overstated merchants' resistance to an increase in the merchant discount and therefore to an increase in the interchange fee. Second, when merchants are allowed to offer cash discounts, a consumer's decision to use a card depends not only on the technological benefit (convenience, theft and fraud control,...), but also on the extra charge for using a payment card. Third, when several payment card systems compete, the opportunity cost for a merchant of accepting a card is endogenous as long as some customers hold cards on multiple systems. For example, a merchant who turns down American Express may see the customer pay with Visa or Master Card rather than with cash or a check. Thus, the earlier literature understated merchants' resistance under system competition.

The paper is organized as follows. Section 2 describes the working of the payment card industry. Section 3 develops the model. Section 4 compares the interchange fee selected by the payment card association with the socially optimal one. Section 5 studies the determinants of merchant resistance. Section 6 discusses the implications of unobserved merchant heterogeneity. Section 7 compares our findings with those in the literature and section 8 summarizes the main insights and discusses some topics for future research. Appendices A to D contain proofs of some of the results.

10 Furthermore, one cannot just set this retail markup to zero by assuming that merchants are undifferentiated Bertrand competitors since the very decision of whether to accept the payment card may be a factor of differentiation among merchants.

2. WORKING OF THE PAYMENT CARD INDUSTRY

A card payment is a service offered to two parties (the cardholder and the merchant) jointly by two other parties (the issuer and the acquirer). Figure 1(a) describes the costs and benefits attached to a card transaction. The total cost of this service is the sum of the issuer's cost c_I and the acquirer's cost c_A. Suppose that the benefit accruing to the cardholder (or buyer) for the marginal use of a payment card is equal to b_B. Similarly, the benefit to the merchant (or seller) of this marginal use of a payment card is b_S. The benefits b_i and costs c_i referred to above are *net* benefits and costs. The cardholder and the merchant must compare the utilities they get by using payment cards with those associated with alternative payment methods (cash, checks,...). At the social optimum, the total benefit of the marginal transaction, $b_B + b_S$, is equal to its total cost, $c_I + c_A \equiv c$. Figure 1(a) also features the payments from end-users to intermediaries: cardholders pay f to consumers and merchants pay merchant discount m to acquirers. These two fees are market determined given the association's choice of interchange fee.

The key feature of payment systems, and one that arises in several other industries characterized by network externalities (media, software, matchmakers, etc.), is its two-sidedness.

Whether the transactions occur within a cooperative undertaking as studied here, or through a for-profit company (such as Amex) playing both roles of issuer and acquirer, the system must attract both sides of the market. Any contemplated increase in the merchant discount must carefully consider the likely merchant resistance; and similarly on the cardholder side.

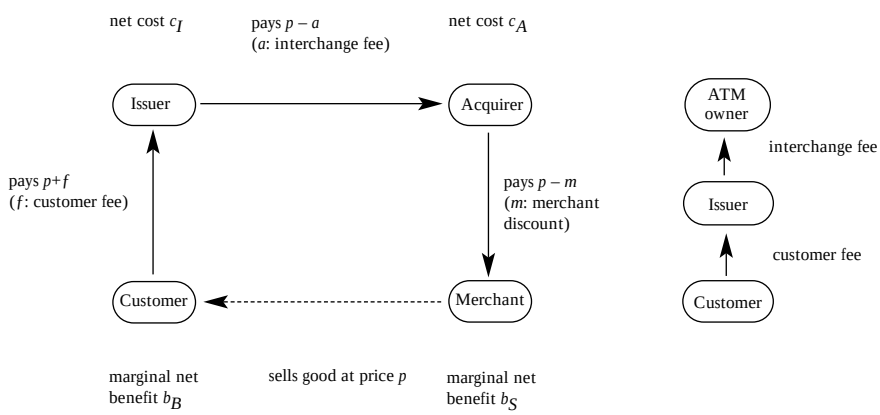

Figure 1

In that respect, the card payment industry is fundamentally different from ATM networks (see figure 1(b)). There is no counterpart to merchant resistance in ATM markets, which are one-sided markets (there is only one side in the market to attract).[11] This one-sidedness has several consequences. First, a change in the interchange fee (the fee paid by the customer's bank to the ATM owner in this case) does not affect another side of the market with whom the customers enjoy network externalities. Second, the issuers' decision to raise the interchange fee raises issuer marginal cost and softens competition in the ATM case; it lowers their marginal cost of serving customers and thereby enhances issuer competition in the card payment context. Third, the choice of the interchange fee allows issuers to tax each other in the ATM context, but does not do so for card payments.[12] Last, the primary role of the interchange fee is to set the price level in the ATM context, and to alter the relative price structure between the two sides of the credit card market, as we will argue. The two industries are therefore markedly different.

3. A MODEL OF THE PAYMENT CARD INDUSTRY UNDER THE NO-SURCHARGE RULE

In our basic model, there is a single payment card association. Except in section 5.4, we assume that the payment card association prohibits merchants from imposing surcharges on customers paying with a card (as Visa and MasterCard currently do). Section 5.4 looks at the impact of preventing the card association from adopting this policy. Our analysis makes two simplifying assumptions. Although both can easily be relaxed, these two assumptions fit well the payment card industry. First, we assume that acquirers are competitive while issuers have market power. The acquiring side involves little product differentiation as well as low search costs and is widely viewed as highly competitive.[13] In contrast, the issuing side is generally regarded as exhibiting market power. The cause and the extent of market power is highly country-specific. It may be due to innovation[14] or to other factors such as search costs, reputation, or the nature of the card.[15] In the model below, we assume that issuers have some market power. Note that, were the issuing side perfectly competitive,

11 Another difference is that ATM networks are three—rather than four—party networks. But that difference is not crucial; indeed, the same issues as those analyzed in this paper arise in for-profit payment systems, which involve only three parties (the system, the customer and the merchant).

12 Antitrust dimensions of ATM networks are studied in Baker (1995), while McAndrews (1996) develops a model of fee-setting for ATM network services.

13 See, e.g., Evans-Schmalensee (1999, chapter 6).

14 Attractive frequent user programs, payment facilities, co-branding and single bill offerings (telephone and payment card for example), corporate card, and so on.

15 In France, payment cards are primarily debit cards; the payment is automatically debited at the end of the month from the customer's bank account (thus credit is limited to intra-month credit). Consumers therefore use debit cards issued by their banks. Yearly fees are high.

issuers would have no preference over (make no profit regardless of) the interchange fee, and so the latter would be indeterminate.[16] The second simplifying assumption is that customers have a fixed volume of transactions, normalized to one transaction.[17] The absence of uncertainty about the number of transactions implies that, from the point of view of the issuer-customer relationship, there is no difference between a fixed yearly fee and a per transaction customer fee.[18] There is endogeneity of the volume of *payment card* transactions, though, because consumers may choose not to have a card, or may be unable to use their card if the merchant refuses it.

Following Section 2, let c_I and c_A denote the per transaction cost, incurred by an issuer and an acquirer, respectively. The interchange fee is denoted a and the merchant discount m. The customer's yearly fee is equal to f and for the moment there is no variable (per actual transaction) fee.[19] Last, let b_B and b_S denote the customer's and the merchant's (per transaction) benefit from using the card rather than an alternative payment method, say cash. In the basic version of our model, all merchants have the same benefit while consumers are heterogenous. We will later allow for merchant heterogeneity.

Consumers: Consumers differ as to their benefit from using a payment card rather than an alternative payment method. For example, some customers have an easy access to cash or a low value of time of going to get cash before shopping, while others attach a high value to the convenience afforded by the use of cards. The benefit b_B is continuously distributed on an interval

16 As noted by Schmalensee (2001).

17 The assumption of an inelastic demand for retail goods, which is reasonable in a first step analysis, implies that the interchange fee impacts only the diffusion of payment cards. While a higher interchange fee raises retail prices, when demand is inelastic this increase in retail prices only amounts to a redistribution of surplus between non-cardholders and cardholders, and is neutral from the point of view of aggregate surplus. On the other hand, with a downward sloping demand, changes in retail prices also affect final demand and thus aggregate surplus. However, the global impact of a higher interchange fee on final demand is ambiguous, because a greater diffusion of cards has also a positive impact on the demand for retail goods (as new cardholders buy more) which may offset the negative impact due to the retail price increase.

18 The decomposition of the payment between the two elements however matters for the association: see Section 5.1.

19 The analysis of optimal price discrimination and volume discounts offered by issuers is interesting in its own right, but it is somewhat orthogonal to the problem at hand; the fixed number of transactions allows us to ignore it. Again, see section 5.1 for the implications of variable fees for the association.

$\left[\underline{b}_B, \bar{b}_B\right]$. The fraction of consumers with benefit less than b_B is given by the cumulative distribution function $H(b_B)$, with density $h(b_B)$. The hazard rate $h/(1-H)$ is increasing, in order to guarantee concavity of the optimization programs. Let

$$E\left(b_B | b_B \geq b_B^*\right) \equiv \frac{\int_{b_B^*}^{\bar{b}_B} b_B h(b_B) db_B}{1 - H\left(b_B^*\right)}$$

denote the expected benefit enjoyed by an average *cardholder* (as opposed to consumers in general) when consumers with type $b_B \geq b_B^*$ purchase the card, and those with type $b_B < b_B^*$ do not.

Issuers: Each issuer has market power over its customers. We further assume in a first step that issuers are not in the acquiring business. Appendix A shows that, due to the competitiveness of the acquiring business, issuers are actually indifferent between entering the acquiring business and staying out at the equilibrium interchange fee (and so they may actually be in the acquiring business after all), and furthermore that they would not benefit from an interchange fee that creates a strict preference for them to enter the acquiring segment.

Assuming that the card is accepted by all merchants (an aspect which we will need to investigate in our equilibrium model), a customer with benefit b_B and facing customer fee f purchases the card if and only if

$$b_B \geq f.$$

For expositional simplicity, let us focus on a symmetric oligopolistic equilibrium, in which all issuers in equilibrium charge the same customer fee f. Let $D(f)$ denote the total demand for cards, and $\beta(f)$ the average card holder benefit. That is

$$D(f) \equiv 1 - H(f),$$

and

$$\beta(f) \equiv E\left[b_B | b_B \geq f\right].$$

Note that the demand for cards, $D(f)$, decreases with the customer fee, and that the average cardholder benefit, $\beta(f)$, is increasing and bounded.

The net cost of a transaction for an issuer is equal to the difference between the "technological cost" c_I and the interchange fee a.[20] Let $f = f^*(c_I - a)$ denote the equilibrium customer fee. We make the following regularity assumption:

Assumption 1: The oligopolistic equilibrium fee, $f^*(c_I - a)$, is defined for all values of the interchange fee a $(\text{even } a > c_I)$ and decreases with it. Each member bank's profit increases with the interchange fee a.

To understand Assumption 1 note that there are two competing economic forces associated with a change in an industry's marginal cost. Take an industry with N symmetrically differentiated firms with marginal cost c. Let $p^*(c)$ denote the oligopolistic equilibrium price and $D(p)$ denote the total demand when all prices are equal to p. The industry's total profit, equal to N times the per-firm profit is

$$\pi(c) = \left[p^*(c) - c\right] D\left(p^*(c)\right).$$

An increase in industry marginal cost has a direct negative impact on profitability (it decreases the per transaction revenue), as well as an indirect impact through the price change (competition becomes less intense). If $0 \leq dp^*/dc \leq 1$, then the direct effect dominates the indirect one (there is some cost absorption) and Assumption 1 is satisfied. We think that Assumption 1 is a mild assumption on empirical grounds. Industry associations never lobby for an increase in the prices of their inputs (labor, intermediate inputs, etc.) or for an increase in ad valorem taxes levied on their industry. Theoretical considerations also provide substantial support for the hypothesis. While Assumption 1 only holds weakly in the case of perfect competition with constant returns to scale (as there is no profit), it holds strictly (the profit strictly increases with the interchange fee) in standard models of oligopolistic competition; let us provide a few examples satisfying Assumption 1:

Example 1: Monopoly issuer. A monopolist chooses its fee so as to maximize $\left[f + (a - c_I)\right]\left[1 - H(f)\right]$. A simple revealed preference argument shows that this fee is a decreasing function of the interchange fee. That is, a monopoly issuer finds it more costly to restrict the number of payment cards and to exercise its market power if the interchange fee increases.[21] Moreover, from the envelope theorem, the issuer's profit decreases with its net cost, and therefore increases with the interchange fee.

---

20 This difference may be negative. A negative issuer marginal cost (which arises when $a > c_I$) does not create the problem of unbounded consumption usually associated with negative marginal costs; for, even if the customer is offered cash-back bonuses or other inducements to use the card (see section 5.1), the customer must still pay for the merchant's good, that is cannot use the card "on a stand-alone basis".

21 The reader will here recognize the standard argument that a proportional subsidy to firms with market power reduces the distortion due to excessive margins.

Example 2: Symmetric Cournot oligopoly. Assumption 1 is satisfied in a symmetric Cournot oligopoly whenever the elasticity of demand exceeds one (Seade 1987).

Example 3: Hotelling model with outside goods. In the classic Hotelling model with a covered market, the price charged by the firms reacts one-for-one with the cost, and so profits are invariant to the interchange fee. With outside goods (alternative means of payment here), profits strictly increase with the interchange fee, as the lower price allows issuers to gain market share from these alternative means of payment.

An analysis focused on the issuing side is incomplete. To understand the impact of the interchange fee, we must perform an equilibrium analysis. For, the interchange fee also impacts the merchant discount, and therefore the merchants' willingness to accept the card. In turn, the customers' willingness to purchase a card depends on the number of merchants accepting it. Last, prices charged by merchants to customers may depend on the interchange fee.

Acquirers: Acquirers face per transaction cost c_A and are competitive. Thus, for interchange fee a, they offer merchant discount m given by

$$m = a + c_A. \tag{1}$$

Because they are competitive, acquirers play no role in our analysis. They just pass through the interchange fee to the merchants.[22]

Merchants: To study the impact of the interchange fee on final prices and social welfare, we use the standard Hotelling model of the "linear city"[23] (or cities: there may be an arbitrary number of such segments). Consumers are located uniformly along a segment of length equal to 1. Density is unitary along this segment. There are two stores selling the same physical good and located at the two extremes of the segment. Consumers wish to buy one unit and for this transaction must pick a store. They incur transportation cost t per unit of distance. As is usual, this transportation cost is meant to reflect the facts that

22 Acquirers with market power would care about the interchange fee, and so one would need to consider, as Schmalensee (2001) does, the relative strength of issuers and acquirers within the payment card associations. To some extent, the two groups have conflicting interests with acquirers in favor of an interchange fee lower than the issuers' preferred level. Also, there is now some incentive for vertical integration, so as to limit the double marginalization in the provision of payment card services. Furthermore, and from a social point of view, the interchange fee must now reduce two distortions: it must be high in order to subsidize issuers and low so as to subsidize acquirers; a single instrument cannot achieve these two conflicting goals. Finally, as Schmalensee (2001) emphasizes, providing proper incentives to both sides in this "moral-hazard-in-teams" problem would require outside funding at the margin.

23 See, e.g., Tirole (1988).

products or services are differentiated and that different consumers prefer different products. Let d denote each firm's unit manufacturing/marketing cost (gross of the merchant discount). We normalize d so that it includes transaction costs associated with cash payments. Merchants enjoy benefit b_S per payment card transaction.[24] We assume that

$$\bar{b}_B + b_S > c_I + c_A. \tag{2}$$

If condition (2) were violated, payment cards would generate no social surplus.

Merchants $i = 1,2$ set their retail prices (p_1, p_2) noncooperatively as in Hotelling's model.They also decide noncooperatively whether to accept payment cards. We assume that the two decisions are sequential: card acceptance is followed by price setting (this is not crucial). Last, for the sake of conciseness, we will focus on "interior solutions". That is, a merchant never corners all consumers of a given type even if he is the only merchant to accept payment cards.[25]

Determination of the interchange fee: We will consider the two cases in which the issuers and a social planner maximizing total surplus, respectively, choose the interchange fee. Acquirers are indifferent as to the level of this fee.

Timing: The timing is as follows:

Stage 1: The interchange fee is set (either by the issuers or by a central planner).

Stage 2: Issuers set fees for their customers, who elect or not to have a card. Merchants decide whether to accept payment cards, and then set their retail prices.

Stage 3: Customers observe the retail prices and whether cards are accepted, and pick a store. If the selected store does not accept payment cards or if the consumer does not own a payment card, the consumer must incur his opportunity cost (b_B) of using the alternative payment method; and similarly the merchant incurs opportunity cost b_S.

24 At this stage we assume homogeneity of merchants. This assumption is relaxed both in the remark after Proposition 1 (merchant observed heterogeneity) and in Section 6, where we allow for merchant unobserved heterogeneity.

25 This assumption requires that $\bar{b}_B$ not be too large relative to t.

4. SOCIALLY AND PRIVATELY OPTIMAL INTERCHANGE FEES

4.1 Merchant behavior

Let us now analyze the model described in section 3. Let us for the moment take the interchange fee as given. Because $f^*(c_I - a)$, the equilibrium customer fee in the oligopolistic issuing market, is a decreasing function of the interchange fee, the average benefit of a *cardholder*, $\beta[f^*(c_I - a)]$ is decreasing in a: the higher the interchange fee, the lower the customer fee; and so customers with lower willingnesses to pay for a card are induced to take a card when the interchange fee increases.

Let
$$m^n(a) \equiv m - b_S = c_A + a - b_S \tag{3}$$

denote the net cost (merchant discount minus merchant's benefit) for a merchant of selling to a cardholder rather than to a consumer using an alternative payment method. Note that this net cost does not embody possible strategic effects of accepting cards in the merchant's competitive environment. Finally, let $\bar{a}$ be uniquely[26] defined by

$$\beta[f^*(c_I - \bar{a})] = m^n(\bar{a}). \tag{4}$$

In words, $\bar{a}$ *is the level of the interchange fee at which the net cost to the merchants is equal to the average cardholder benefit*.

Last, in order to analyze how merchant resistance is affected by the diffusion of cards in the population, let us parameterize the oligopolistic equilibrium fee by a number τ that increases when competition among issuers becomes more intense: $f = f^*(c_I - a, \tau)$, with f^* decreasing in τ.

Proposition 1 *(i) Under the no-surcharge rule, there exists an equilibrium in which all merchants accept the card if and only if* $a \leq \bar{a}$.

(ii) As competition among issuers intensifies, merchant resistance increases, i.e., the maximal interchange fee $\bar{a}$ *decreases*.

26 The left-hand side of (4) is decreasing in $\bar{a}$, while the right-hand side is increasing; so there is at most one solution. To prove existence, note that the left-hand side of (4) is bounded, while the right-hand side can take arbitrarily small and arbitrarily large values.

Proof of Proposition 1

(i) Suppose that consumers expect merchants to accept the card. Then issuers charge $f^*(c_I - a)$, and the demand for cards is $D\left(f^*(c_I - a)\right)$. Is it indeed optimal for all merchants to accept the card? Suppose they do. Then a merchant's average cost per customer is $d + D\left(f^*(c_I - a)\right)\left[c_A + a - b_S\right] = d + D\left(f^*(c_I - a)\right)m^n(a)$. As is usual in the symmetric Hotelling model, the equilibrium price p^* is the same for both merchants and is equal to the merchants' marginal cost plus the transportation cost:

$$p^* = \left[d + D\left(f^*(c_I - a)\right)m^n(a)\right] + t. \tag{5}$$

Each merchant's profit is equal to the margin times the market share:

$$\pi^* = \frac{t}{2}. \tag{6}$$

To show this, note that, for given prices $(p_i,\ p_j)$, merchant i's market share x_i among customers of type b_B is independent of b_B (since a customer pays either cash or with a payment card, independently of the merchant) and is given by $p_i + tx_i = p_j + t(1 - x_i)$, yielding

$$x_i = \frac{1}{2} + \frac{p_j - p_i}{2t}. \tag{7}$$

So, merchant i solves

$$\max_{p_i}\left\{\left[p_i - \left(d + D\left(f^*(c_I - a)\right)m^n(a)\right)\right]x_i\right\},$$

yielding, at equilibrium, equations (5) and (6).

Suppose now that merchant i deviates from this presumed equilibrium by not taking the card. Consumers with type $b_B < f = f^*(c_I - a)$ don't have a card, and are not affected by merchant i's decision. So, merchant i's market share among these customers is still given by (7). In contrast, merchant i's market share is reduced (for a given price) among cardholders. Among cardholders with benefit b_B, this market share is given by $p_i + tx_i = p_j + t(1 - x_i) - b_B$,

or

$$x_i(b_B) = \frac{1}{2} + \frac{p_j - p_i - b_B}{2t}. \tag{8}$$

Aggregating over all customers (cardholders and non cardholders), merchant i's market share is

$$x_i = \frac{1}{2} + \frac{p_j - p_i - D(f)\beta(f)}{2t}.$$

On the other hand, merchant i's margin has increased to $(p_i - d)$. So, merchant i solves

$$\max_{p_i} \{(p_i - d)x_i\},$$

yielding price

$$p_i = \frac{1}{2}[p_j + t + d - D(f)\beta(f)]. \tag{9}$$

On the other hand, the composition of merchant j's market share has changed, since the proportion of cardholders has increased. His profit function becomes:

(10)

$$\pi_j = [1 - D(f)](p_j - d)\left(\frac{1}{2} + \frac{p_i - p_j}{2t}\right) + D(f)(p_j - d - m^n)\left(\frac{1}{2} + \frac{p_i - p_j + \beta(f)}{2t}\right).$$

Merchant j's optimal price is therefore:

$$p_j = \frac{1}{2}[p_i + t + d + D(f)(\beta(f) + m^n)]. \tag{11}$$

The equilibrium prices are obtained by solving the system {(9), (11)}:

$$p_i = t + d - \frac{1}{3}D(f)[\beta(f) - m^n], \tag{12}$$

$$p_j = t + d + \frac{1}{3}D(f)[\beta(f) + 2m^n]. \tag{13}$$

When merchant i decides to refuse the card, this increases not only the market share of merchant j (for given prices p_i, p_j) but also the average cost of merchant j (see (10)), since the proportion of cardholders in his clientele increases. For both reasons, merchant j's equilibrium price increases (see (13)). The "high-quality merchant", merchant j, charges a higher price. Equilibrium profits are given by:

$$2t\pi_i = \left[t - \frac{1}{3}D(f)(\beta(f) - m^n)\right]^2 \tag{14}$$

$$2t\pi_j = \left[t + \frac{1}{3}D(f)(\beta(f) - m^n)\right]^2 - m^n\beta(f)D(f)(1 - D(f)). \tag{15}$$

Comparing with (6), we see that the deviation is profitable $(\pi_i > t/2)$ if and only if:

$$\beta(f) < m^n.$$

Since $\beta(f)$ is decreasing in a, while m^n is increasing in a, we see that universal acceptance is an equilibrium if and only if $a \leq \bar{a}$, which ends the proof of (i).

(ii) The maximum interchange fee that merchants accept is now a function $\bar{a}(\tau)$, implicitly determined by the relation:

$$\bar{a}(\tau) + c_A = b_S + \beta\left[f^*\left(c_I - \bar{a}(\tau), \tau\right)\right].$$

Since β is increasing, and f^* is also increasing in its first argument, $\bar{a}$ decreases in τ. As competition among issuers becomes more intense, merchant resistance increases and the interchange fee $\bar{a}$ and the customer fee f decrease.[27] Issuer competition makes the card available to a wider clientele and thereby lowers the average cardholder's benefit. Merchants are then less inclined to take the card.

Remark on observed merchant heterogeneity: Proposition 1 can be straightforwardly generalized to allow for observable merchant heterogeneity: see Appendix B. For example, in the early 90s,Visa lowered its interchange fee for supermarkets, which have a lower demand for credit card services than other shops. Assuming that the issuers do not charge cardholders on the basis of which category of merchants they patronize, merchant heterogeneity generates an externality among merchants: An issuer marketing a card internalizes the *average* interchange fee that he will receive when the consumer will make card transactions at different kinds of stores. So for example, when supermarkets are brought into the system through a low supermarket-specific interchange fee, the average interchange fee decreases and so the cost to issuers increases. Appendix B shows that the individual interchange fee for all other categories of merchants increases. Intuitively, cards become more expensive, which raises the average cardholder benefit and reduces merchants' resistance.

27 Indeed the implicit function theorem gives:

$$\frac{d\bar{a}}{d\tau} = \frac{\beta' f_2^*}{1 + \beta' f_1^*} < 0,$$

where lower indices denotes partial derivatives. Similarly we can compute the impact on the equilibrium customer fee:

$$\frac{d}{d\tau}\left[f^*\left(c_I - \bar{a}(\tau), \tau\right)\right] = \frac{f_2^*}{1 + \beta' f_1^*} < 0.$$

The proof of Proposition 1 illustrates the externalities involved in the acceptance decision. A potential multiplicity (stated in Proposition 2 and proved in Appendix C arises because one merchant's decision to reject the card raises the other merchant's average cost of serving a customer and therefore makes the latter more reluctant to accept the card:

Proposition 2 *The merchants' card acceptance policies exhibit strategic complementarity.*

For $a > \bar{a}$*, card rejection is the unique equilibrium.*

For $a \leq \bar{a}$*, either both merchants accept the card or both refuse it. For a equal to or slightly below* $\bar{a}$*, both merchants' refusing the card is also an equilibrium. The rejection equilibrium is less likely to exist when merchant differentiation increases (that is, if it exists for* (a,t)*, then it also exists for* (a,t') *with* $t' < t$*. The set of interchange fees such that the rejection equilibrium exists is included in* $(\underline{a},\infty)$ *where* $\underline{a} \equiv b_S - c_A$.

4.2 Determination of the interchange fee

From Proposition 2, the "merchant acceptance subgame" admits either a unique equilibrium or two (pure strategy) equilibria, depending on the level of the interchange fee. For $a > \bar{a}$, the card is always rejected. For $a \leq \bar{a}$, there exists a "low-resistance equilibrium" in which merchants accept the card; there may further exist a "high-resistance equilibrium" in which merchants reject the card. When the two equilibria coexist, issuers prefer the low-resistance equilibrium. For expositional simplicity, we will focus on the low-resistance equilibrium described in Proposition 1. But it will be clear that our welfare conclusions (presented in Proposition 3 below) hold *for any equilibrium selection* as long as $\bar{a}$ is replaced more generally by "the highest interchange fee $\hat{a}(\leq \bar{a})$ that induces merchant acceptance".

Because the issuers' profit is increasing in a, the optimal interchange fee for the issuers is the highest level that is consistent with the merchants' accepting the card, namely $a = \bar{a}$, corresponding to a customer fee

$$f = f^*(c_I - \bar{a}). \tag{16}$$

Suppose instead that a benevolent and omniscient social planner selects the interchange fee so as to maximize total welfare

$$W(f) = [\beta(f) + b_S - c_A - c_I]D(f) = \int_f^{\bar{b}_B} [b_B + b_S - c_A - c_I]dH(b_B).$$

Ignoring in a first step the constraint that the merchants must accept the card,

at the socially optimal interchange fee a, the total cost and benefit of the marginal transaction are equal, or

$$f = f^*(c_I - a) = c_I + c_A - b_S.$$

We are thus led to consider two cases:

$$\text{(i) } c_I + c_A - b_S \leq f^*(c_I - \bar{a}).$$

In this case, the socially optimal provision of payment cards requires a low customer fee, which can be obtained only through an interchange fee that exceeds the level at which merchants accept the card. The socially optimal interchange fee is then equal to $\bar{a}$ and thus coincides with the issuers' preferred interchange fee.

$$\text{(ii) } c_I + c_A - b_S > f^*(c_I - \bar{a}).$$

In this case, the socially optimal interchange fee is smaller than the issuers' preferred interchange fee. This means that a payment card association controlled by issuers selects an interchange fee that leads to an *overprovision*[28] of payment card services.

Proposition 3 *Under the no-surcharge rule, the issuers' preferred interchange fee is equal to* $\bar{a}$.

(i) If $c_I + c_A - b_S \leq f^*(c_I - \bar{a})$, *then the socially optimal interchange fee is equal to the issuers' preferred interchange fee.*

(ii) If $c_I + c_A - b_S > f^*(c_I - \bar{a})$, *the interchange fee set by a payment card association controlled by issuers leads to an overprovision of payment card services.*

It is easily verified that for example with a monopoly issuer facing a linear demand each of the two cases envisioned in Proposition 3 may arise. Also, and as noted above, Proposition 3 would continue to hold if we presumed a higher degree of merchant resistance; the only difference is that overprovision of card services would become less likely: a low merchant resistance is the worst case scenario for the social optimality of an issuer-determined interchange fee.

28 This overprovision of card services is analogous to the overprovision of payment services by checks that occurs in some countries where regulation prevents banks from charging customers for the use of checks. The analysis above assumes that competing means of payment such as checks are not distorted. The analysis of the socially optimal interchange fee in the presence of subsidized checks is a straightforward adaptation of that above.

5. FOUR DETERMINANTS OF MERCHANT RESISTANCE

A key factor in both the positive and the normative analyses is the degree of merchant resistance. Lower merchant resistance is more conducive to overprovision, a higher resistance to social optimality of the association determined fee. The analysis of the low-resistance equilibrium described in Proposition 1 unveils the reason why overprovision may occur: Merchants, when deciding whether to take the card, consider the convenience benefit of the *average* cardholder rather than the (lower) benefit of the *marginal* cardholder as the welfare analysis would command. We now discuss factors impacting merchant resistance to payment cards and therefore the likelihood that the card association under- or over-provides card payment services.

5.1 A basic externality in the issuers' choice of pricing structure.

We noted that it does not matter from the point of view of the issuer-customer relationship whether the issuer charges a fixed yearly fee or a volume proportional fee.[29] Interestingly, this "irrelevant" contractual choice turns out to make a big difference collectively, since it affects merchant resistance.

To see why, let us replace our assumption that the issuers offer yearly fees with no proportional payments by the opposite polar assumption[30] of a "perfect two-part tariff" with marginal cost pricing for the variable part; that is, for a number n (that we have normalized to one) of transactions and yearly fee F_i charged by issuer i, the total charge for the cardholder is:

$$f_i = F_i + (c_I - a)n = F_i + (c - m)n.$$

The analysis of competitive edge of card acceptance summarized in Proposition 1 carries over, as long as the convenience benefit b_B of using the card is replaced by the net benefit $b_B - (c - m)$ for the consumer of a card transaction. Competition among issuers for a given interchange fee a and universal card acceptance is unchanged: The equilibrium (now total) fee is still $f^*(c - m)$, and so the equilibrium (symmetric) yearly fee is defined as the residual

$$F^* = f^*(c - m) - (c - m).$$

29 Recently, many Visa and MasterCard banks have introduced no-fee offerings. In contrast, Visa and MasterCard cards in Europe carry substantial yearly fees.

30 More general models of price discrimination usually exhibit tariffs that are intermediate between a volume-insensitive fee and a perfect two-part tariff. The basic point made here—the externality in the choice of pricing structure—would still hold.

The highest interchange fee $\hat{a}$, or equivalently the highest merchant discount $\hat{m} = \hat{a} + c_A$, that is consistent with universal card acceptance is therefore given by

$$\hat{m} = b_S + E\left[b_B - (c - \hat{m}) \,|\, b_B - (c - \hat{m}) \geq f^*(c - \hat{m})\right] \tag{17}$$

instead of (4), which can be rewritten as:

$$\overline{m} = b_S + E\left[b_B | b_B \geq f^*(c - \overline{m})\right]. \tag{18}$$

We can now state the analog to Proposition 1 for variable user payments:[31]

Proposition 4 *When issuers use perfect two part tariffs, there exists an equilibrium in which all merchants accept the card if and only if* $a \leq \hat{a}$. *Furthermore* $\hat{a} > \overline{a}$ *if and only if* $\overline{a} > c_I$.

Thus, variable payments reduce merchant resistance if and only if the interchange fee exceeds the issuer cost. When the interchange fee exceeds the issuer cost, variable pricing rewards the cardholder for using the card; the cardholder is then even more upset when a shop turns down the card, as she loses the reward on top of the convenience benefit. This of course is more than a theoretical possibility. Many Visa and MasterCard banks as well as proprietary cards have introduced inducements for customers to use their card: cashback bonuses (Discover), discounts on products sold by affiliates, travel insurance, frequent flyer mileage, and so forth. In the case of associations, the noncooperative introduction of these volume related payments creates a positive externality among issuers, an externality that is fully internalized in the case of a proprietary system.

5.2 Consumer information

We have assumed that customers are always fully informed about individual merchants' acceptance policy. However, some consumers may not know which stores accept the card and may not leave the store once they patronize it and learn that it does not accept the card. Let us therefore consider the more general case where only a proportion $\alpha \leq 1$ of customers are informed of which merchants accept the card before they select a store (or equivalently customers are informed for only a fraction α of their purchases). For simplicity, let

31 Only the last sentence requires a proof. To derive it, consider the two functions of *m* defined by the left-hand side minus the right-hand side of equations (17) and (18), respectively. These two functions are increasing in *m*, are equal for $m = c$, and that built from (18) is bigger for $m = c$, and smaller for $m > c$.

us further assume that the consumer either is informed of the acceptance policy of *both* merchants in the market for the good he considers buying (and this with probability α) or is uninformed of their acceptance policy (with probability $1-\alpha$). The condition for universal acceptance to be an equilibrium becomes:

$$m^n(a) \leq \alpha\beta\left[f^*(c_I - a)\right],$$

which implies that the maximum possible interchange fee (given by this condition satisfied with equality) is smaller than $\bar{a}$.[32] Our analysis and Propositions 1 through 3 are otherwise unchanged.

Proposition 5 *Suppose that the association-determined interchange fee induces an overprovision of card services under consumer full information about card acceptance* $(\alpha = 1)$. *Then, there exists* $\alpha^* \in (0,1)$ *such that the association determined interchange fee is socially optimal if and only if* $\alpha \leq \alpha^*$. *If the association determined interchange fee is socially optimal for* $\alpha = 1$, *then it is also socially optimal for any* α.

5.3 System competition

We consider now a situation in which two associations $(i = 1,2)$ compete for offering payment card services to customers and merchants. We denote by $a_i (i = 1,2)$ the interchange fees chosen by the associations, by $f_i (i = 1,2)$ the customer fees and by $m_i (i = 1,2)$ the merchant discounts.

We will not attempt to provide an in-depth analysis of system competition here. We however make two points that demonstrate that intuitions based on competition between for-profit corporations are misleading when applied to associations, and thereby stress the need for further research. We show, first, that *competition between two associations need not result in a lower interchange fee, and, second, that even if it does lower the interchange fee, this reduction may lower welfare.*

We maintain the assumption that acquirers are competitive: $m_i = c_A + a_i (i = 1,2)$. Imperfect competition between issuers within each association in general becomes more complicated to model since it is in general influenced by the interchange fee charged by the competing association.[33] Let us therefore look at two simple cases. Suppose first that each customer holds at

32 While consumers may not be aware of whether the two merchants in a particular market take the cards, they have rational expectations (or are informed by the payment card association) as to the fraction (here 1) of merchants accepting the card.

33 Also, we abstract from the difficulties created by duality, *i.e.* the fact that issuers typically belong to both associations (see Hausman et al., 1999).

most one card. Then system competition has no impact on merchant resistance and the analysis of Section 4 is unchanged: both associations choose the maximum interchange fee that is compatible with merchants' acceptance:

$$a_1 = a_2 = \bar{a}.$$

This is because the incentives of the associations and of the cardholders are perfectly aligned: both want to maximize the interchange fee under the constraint that the card is accepted by the merchants.

Second, let us assume that at least some consumers hold two cards not on the same system. While $(\bar{a}, \bar{a})$ is the equilibrium when consumers hold a single card, it in general is not an equilibrium here. Suppose that system 1 "undercuts" and chooses a slightly lower interchange fee. Then merchants, who for interchange fees $(\bar{a}, \bar{a})$ are indifferent (individually) between accepting a given card and rejecting it, now prefer to reject system 2's card, since the consumer may have the other card in her wallet and this card carries a lower merchant discount. *System competition increases merchant resistance*. Note last that from Proposition 3, $(\bar{a}, \bar{a})$ may lead to the socially optimal allocation. Thus, system competition may reduce social welfare by lowering the interchange fee. We of course do not want to draw general welfare implications from this special case,[34] and only want to warn against "natural conclusions" and to stimulate further research on this very interesting topic.

5.4 Cash discounts and the no-surcharge rule

Let us now investigate the implications of lifting the no-surcharge rule. For concreteness, let "cash" be the alternative method of payment. We ignore the transaction costs associated with merchants' charging two different prices. Even so, allowing card surcharges has ambiguous welfare consequences. In essence, cash discounts raise the cost of payment cards and lead to a suboptimal diffusion of that means of payment.

When merchants are allowed to apply card surcharges, their accepting the card is no longer an issue, since they can charge a price for payment card transactions at least equal to the cash price plus their cost of payment card transactions.[35]

34 The assumptions are very strong. Furthermore, it is assumed exogenously that some consumers hold two cards. This decision is endogenized in our follow-up paper (Rochet-Tirole 2001).

35 In our model, merchants face no fixed cost of accepting payment cards. If they did (and that cost were not subsidized by payment card associations), then they might refuse payment cards for a high enough merchant discount.

Proposition 6 *In the absence of transaction costs associated with the merchants' charging different prices:*

(i) For a given interchange fee, allowing card surcharges raises the merchant price for cardholders and lowers it for noncardholders.

(ii) When the no-surcharge rule is lifted, the interchange fee is neutral and there is an underprovision of card services.

(iii) Lifting the no-surcharge rule reduces social welfare in case (i) of Proposition 3. Lifting the no-surcharge rule may increase or reduce social welfare in case (ii) of Proposition 3.

The intuition behind Proposition 6 is as follows. In the absence of transaction costs, merchants charge a higher price to cardholders. When the merchants compete a la Hotelling, they just pass through to the cardholder the increase in their net cost $(m - b_S)$ due to a card transaction. This pass through of the interchange fee by merchants to cardholders prevents the interchange fee from affecting card diffusion implying the neutrality of the interchange fee.

Merchant price discrimination reduces the demand for payment cards. Issuers focus on the high end of the market and no longer attract consumers who are not willing to pay much in the first place and who know that they will face a second markup when paying with the card in the store. Put differently, the card surcharge in stores raises the issuers' cost of providing cardholders with a given surplus of using the card and thus inhibits the diffusion of cards.

In case of initial underprovision of cards (case (i) of Proposition 3), the card surcharge aggravates this underprovision and therefore reduces welfare. In case of overprovision (case (ii) of Proposition 3), the card surcharge offers a countervailing force and may result in a welfare increase.

6. UNOBSERVABLE MERCHANT HETEROGENEITY AND BUSINESS MODELS.

Yet another source of merchant resistance is unobservable merchant heterogeneity, which prevents a payment system from extracting the merchant's individual willingness to pay. Let us derive a few insights concerning its consequences. If b_S is a random variable distributed according to a cumulative distribution function K, the acceptance decision by merchants becomes elastic. For example with uninformed consumers $(\alpha = 0$ in section 5.2$)$, the proportion of merchants who accept the card becomes $1 - K(c_A + a)$. This modifies the potential surplus that a customer obtains by holding the card, since he will be able to use it only with probability $1 - K(c_A + a)$. Therefore he will hold the card if and only if:

$$b_B\left(1-K\left(c_A+a\right)\right)\geq f.$$

The total profit of issuers becomes:

$$\pi_I=\left[f+\left(a-c_I\right)\left(1-K\left(c_A+a\right)\right)\right]\left[1-H\left(\frac{f}{1-K\left(c_A+a\right)}\right)\right],$$

or by denoting $b_B^0=\dfrac{f}{1-K\left(c_A+a\right)}$ the valuation of the marginal cardholder,

$$\pi_I=\left(1-K\left(c_A+a\right)\right)\left[b_B^0-c_I+a\right]\left[1-H\left(b_B^0\right)\right].$$

This expression is proportional to the previous expression obtained in formula (1) in the case of homogenous merchants, provided that f is replaced by b_B^0. Let us assume that the statistical distribution of merchants is homogenous across issuers, so that this proportionality result also applies to individual profits of each issuer.

Thus, previous formulas for prices and profits in the imperfect competition game between issuers are modified in a simple way: in particular the equilibrium customer fee in the case of heterogenous merchants is equal to the previous equilibrium fee $f^*\left(c_I-a\right)$ multiplied by the proportion of merchants $1-K\left(c_A+a\right)$ who accept the card. Similarly, the total profit of issuers at equilibrium is equal to the previous one multiplied by $\left[1-K\left(c_A+a\right)\right]$. The issuers' preferred interchange fee depends on the elasticity of the merchants' acceptance function $\left[1-K\left(c_A+a\right)\right]$.

Unobservable merchant heterogeneity allows us to say a few things about the comparison of the business models of an association and of a closed, for-profit system. Note that if merchants all have the same benefit b_S as in section 4, then the for-profit system sets the highest merchant discount, $m=\bar{a}+c_I$, that it can get away with. Hence, there is no difference in merchant discount with an association. The only difference is that the customer fee is in general higher.

A for-profit system does not set an interchange fee properly speaking. However, it does have an implicit interchange fee through the level of the merchant discount. Let us therefore consider a for-profit system that either is vertically integrated (as is American Express today) or offers licenses to banks (as was the case for instance for Bank Americard Service Corporation before the creation of the NBI, now Visa). In our model, because issuers are symmetrical, a two-part-tariff license with a fixed payment and a per-transaction payment to the system is equivalent to vertical integration, provided the system offers a license to all issuers. So, we can just assume that the for-profit system is vertically integrated and sets the customer fee f and the merchant discount m directly.

A key difference between the for-profit and the cooperative paradigms is that the former has two separate instruments and optimizes over the merchant discount and the customer fee, while in the latter the customer fee is determined by issuer competition once the merchant discount/interchange fee is set. In particular, the cooperative must assess the extent to which an increase in the interchange fee is "competed away" through the competition among issuers. A high merchant discount reduces the issuers' marginal cost; if however this marginal cost saving is mostly passed through to the customers, then the issuers may not gain much from the reduction in their marginal cost and should rather choose a low merchant discount to ensure a wide acceptance of the card.

Proposition 7 (whose proof is available upon request) analyzes three standard models of oligopolistic competition among issuers to see whether this intuition is correct:

Proposition 7 *Let m_p and m_c denote, respectively, the merchant discounts chosen (directly) by a proprietary system and (indirectly) by a cooperative of banks.*

(i) When issuer competition is described by the Hotelling model, $m_c < m_p$, under the (weak) assumption that the elasticity of merchant acceptance is small when m is outside the competitive region.[36]

(ii) In the differentiated Bertrand model of issuer competition with linear demands, $m_c < m_p$.[37]

(iii) When issuers compete la Cournot and demands are linear, $m_c = m_p$.[38]

---

36 In *Hotelling's model of product differentiation*, each consumer has a preferred brand and his surplus depends on the "distance" between his own preferred brand and the selected brand's characteristics. Consider (without loss of generality) an Hotelling duopoly in which the two issuers are located at the two extremes of a segment of length one and customers are located uniformly on the segment. If $y(m)$ is the proportion of merchants accepting the card, then the net surplus of a consumer of issuer y located at distance x_i of the issuer is $y(m)[b_B - tx_i] - t'x_i - f_i$, where t and t' are the parameters of volume-related and fixed differentiation, respectively. The proposition assumes that the elasticity of merchant acceptance $\frac{m|y'(m)|}{y(m)}$ is small for low values of m (technically, for merchant discounts such that the issuer industry is not in the Hotelling competitive region). This assumption is mild since one would expect that almost all merchants would accept the card $(y(m) \approx 1)$ for such low merchant discounts.

37 Case (ii) considers linear demands:

$$D_i(f_1, f_2, ..., f_N) = \gamma y(m) - \alpha f_i + \beta\left(\sum_{j \neq i} f_j\right),$$

where α, β and γ are positive, and $y(m)$ reflects the proportion of merchants who accept the card.

38 Cournot competition refers to the competition among undifferentiated issuers, with the numbers of cards as strategic variables.

7. COMPARISON WITH THE LITERATURE

Economic research has only recently started studying the payment card industry. The theoretical and empirical analyses of the US credit card market were initiated by Baxter (1983) and Ausubel (1991), respectively. Similarly, ATM (Automatic Teller Machines) networks have been analyzed only recently by Gilbert (1990), Matutes and Padilla (1994), Baker (1995), McAndrews (1996), McAndrews and Rob (1996) and Kim (1998).[39] As pointed out in Section 2, though, ATM networks do not obey the same economics as payment card networks, though.

The formal literature on access pricing in the payment card industry is meager.[40] The standard reference is Baxter (1983). Baxter confined attention to the competitive case and to passive (as opposed to fully rational) actors. Baxter assumed that merchants accept the card as long as $m^n(a) \geq 0$. That is, merchants believe that accepting the card does not help attracting consumers. This assumption is legitimate provided the consumers are unaware of which stores accept the card and furthermore still buy when they learn that the shop they patronize does not take the card (case $\alpha = 0$ in Proposition 5). Baxter's model overstates merchant resistance by ignoring that card acceptance is a competitive instrument. Baxter also treated consumers as passive actors. In this framework (which cannot predict the choice of an interchange fee by an association), Baxter performed the normative analysis of finding the optimal interchange fee.

Schmalensee (2001), in an analysis complementary to ours, analyzes the provision of payment card services as a moral-hazard-in-teams problem. The number of payment card transactions is a function of the issuers' and the acquirers' efforts, with a complementarity between the two efforts.[41] Each side's effort is bidimensional: marketing effort as well as terms given to the banks' clients (merchant discount for acquirers, customer fee for issuers). The Nash equilibrium of the resulting "second stage" game depends on the interchange fee, which is determined in a first stage through bargaining between issuers and acquirers. Schmalensee solves for the outcome of this two-stage game for an arbitrary allocation of bargaining power[42]. Schmalensee argues that there is no support for a public policy of forcing interchange fees to zero.

---

39 McAndrews (1997) studies the impact of the direct presentment regulation, that prevent U.S. banks from charging each other an interchange fee for checks. His results argue in favor of lifting this regulation.

40 Chakravorti and Shah (2001) provide a good survey of theoretical papers on this topic. They also give useful information on common market practices as well as on the regulatory and legal background.

41 Schmalensee first analyzes the case of a monopoly issuer and a monopoly acquirer. He then generalizes the model to oligopolistic competition on both sides.

42 As Schmalensee notes, in the US, banks' voting rights in Visa and MasterCard are more sensitive to issuing volume than to acquiring volume; this suggests that the bargaining power is on the issuing side.

Our paper, like Schmalensee's, analyzes market power issues. It follows Baxter in its emphasis on the determination of the efficient interchange fee; yet, by departing from Baxter's perfectly competitive paradigm and thus from the banks' indifference as to the level of the interchange fee, and by modifying his analysis to account for consumer and merchant rational behavior, our framework allows a comparison between the privately optimal interchange fee (the object of Schmalensee's analysis) and the socially optimal one. Furthermore, and because we derive the demand for payment card transactions from individual consumer preferences and endogenize merchants' demand, we are able to identify the determinants of merchant resistance and to analyze the impact of the no-surcharge rule, which has not yet been studied in the literature.

A number of more recent contributions have built on our framework and extended it in several relevant directions. Wright (2000) provides an in-depth analysis of the distinction between membership and usage decisions. In particular, an interesting hold up problem arises when merchants are monopolies. Suppose that issuers are led to charge a fee for membership, perhaps with a rebate for transactions. When the no surcharge rule (NSR) is lifted, then each monopoly merchant does not internalize the impact of his surcharge on the overall membership decision of the consumer (indeed the fee is already sunk). The merchant's surcharge is set so as to leave no ex post surplus to the cardholder, who therefore does not want to become a cardholder. In contrast, consumers are better protected against hold up in the absence of NSR when, as in our paper, merchants compete with each other.[43] Schwartz and Vincent (2000) investigate another aspect of the NSR. Interestingly, they allow for an elastic demand for goods (at the cost of assuming an inelastic demand for card usage—consumers are split between cardholders and cash users) in a world with a monopoly for-profit payment system. They highlight the impact of the NSR on the double marginalization associated with the interplay between merchant monopoly power and issuer monopoly power. They show that in this environment the NSR generally reduces consumer surplus and often total surplus.[44]

43 Wright also analyzes the interesting possibility that merchants use the card acceptance decision as a differentiation strategy.

44 Chakravorti and Emmons (2000) obtain similar conclusions but for totally different reasons. They use a Diamond-Dybvig type model to capture the impact of credit card pricing on the consumption profile of cash constrained consumers. Since we have focused on the payment activities, this dimension is absent from our paper.

Wright (2001) extends our welfare analysis to merchant heterogeneity (in contrast, our primary goal when introducing merchant heterogeneity was to generate a nontrivial comparison between for-profit and associative business models). One of the main contributions is the comparison between the welfare maximizing interchange fee and the ones that maximize output and banks' profits. Gans and King (2001) substantially generalize the neutrality result in the absence of NSR. They then derive a number of results on the competition between cash only and card accepting merchants and their implications for the determination of interchange fees. Last, in a follow-up paper (Rochet-Tirole 2001), we pay much less attention to the determinants of merchant resistance, and rather provide a general analysis of platform competition. We compare price structures under platform competition and those under a monopoly platform (either for-profit or run by an association) and under a benevolent social planner.

8. SUMMARY AND CONCLUDING REMARKS

To analyze the cooperative determination of the interchange fee, the paper has developed a framework in which banks and merchants may have market power and consumers and merchants decide rationally on whether to buy or accept a payment card. In the absence of unobserved heterogeneity among merchants, an increase in the interchange fee increases the usage of payment cards, as long as the interchange fee does not exceed a threshold level at which merchants no longer accept payment cards. At this threshold level, the net cost for merchants of accepting the card is equal to the average cardholder benefit. The interchange fee selected by the payment card association either is socially optimal or leads to an overprovision of payment card services.

A leitmotiv of our analysis has been the central role played by merchant resistance. A first insight is that, in the absence of unobservable heterogeneity, merchants accept the card even though the merchant discount exceeds the technological and payment guarantee benefit they derive from card acceptance. Payment card systems can exploit each merchant's eagerness to obtain a competitive edge over other merchants. Remarkably, though, the interchange fee need not be excessive. The exploitation of the merchants' search for a competitive edge has two benefits from a social viewpoint: On the merchant side it forces merchants to internalize cardholders' convenience benefit, and on the customer side it offsets the underprovision of cards by issuers with market power. In some circumstances, though, the interchange fee may be too high since merchants' incentives are driven by the average cardholder's convenience benefit rather them the marginal cardholder's.

Merchant resistance is affected by several factors. Better consumer information (obtained through advertising or repeat purchases) about which stores accept the card, or an increased consumer willingness to quit the store when discovering it does not accept the card (due to the size of the payment or the

proximity of a similar store) lower merchant resistance. Cash-back bonuses or other inducements offered by issuers for card usage also weaken merchant resistance. We would therefore expect associations not to mind when their members offer such inducements (and for-profit systems to make heavy use of these inducements), while in contrast being negatively affected by per-transaction payments charged by issuers.[45] Last, system competition increases merchant resistance when some cardholders have cards on several systems in their pocket.

If the no-surcharge rule is lifted and price discrimination is costless to merchants, the interchange fee no longer impacts the level of payment card services. The merchant price for cardholders is increased and that for noncardholders decreased. Merchant price discrimination leads to a lower diffusion of card services, whose welfare consequences depend on whether there is overprovision or underprovision of card services under the no surcharge rule.

The paper has focused primarily on associations. However, several insights obtained in this paper carry over to for-profit systems. In particular, the analysis of the various factors impacting merchant resistance is unchanged. Still, it would be worth conducting an in-depth analysis of for-profit systems' strategies.

The payment card industry has received scant theoretical attention, and it won't come as a surprise to the reader that more research is warranted. We argued that the framework developed here can be used as a building block to analyze more general situations with acquirer market power and distorted competing means of payments. The payment card industry offers many other fascinating topics for theoretical and empirical investigation, such as the impact of duality,[46] the governance of payment card associations, the competition between associations and proprietary systems, and the development of E-commerce.

Last, and taking a broader perspective, our analysis initiates the study of markets in which network externalities between multiple sides of the market call for a careful design of the price structure in order to "get all sides on board". Consider the pool containing the patents essential to the implementation of the audio and video MPEG standard. The pool for example sets licensing fees for DVD players and DVD discs.[47] The allocation of the licensing fees between the two sides conditions the speed at which the player and disc manufacturers invest in the MPEG standard and move away from the previous standard, or affects their choice among competing new standards. The player and disc manufacturers are similar to the issuers and acquirers of our model, and consumers and artists resemble cardholders and merchants. More generally, most markets

45 Associations however currently do not prohibit per-transaction payments (which sometimes exist for debit cards).

with network externalities involve multiple sides and the choice of a price structure. Software and videogame platforms must attract developers and users, portals and media advertizers and "eyeballs", real estate agencies buyers and sellers, shopping malls consumers and shops, the Internet websites and consumers, and so forth. The reasons for the nonneutrality of the price structure (and therefore the rationale for a careful design of this structure) is industry contingent and so their analysis of the credit card industry does not directly apply to these other markets. But some of the underlying economics such as the respective elasticity and welfare analysis and the determinants of merchant resistance have much broader applicability than to the credit card industry. We therefore hope that this paper will stimulate new research on these fascinating features of network economics.

---

46 Duality refers to the fact that banks can (and usually do) belong to both Visa and MasterCard. See Hausman et al. (1999) for a start on the analysis of duality.

47 For example, in the agreement approved on June 10, 1999 by the US Department of Justice, royalties were $0.075 per DVD disc and 4% of net sales price of DVD players and decoders with a minimum royalty of $4.00 per player/decoder.

APPENDIX A:

Absence of benefit from vertical integration

If the acquiring business is competitive, there is no strict incentive for an issuer to integrate with an acquirer. Suppose indeed that an issuer merges with an acquirer (or enters the acquiring business) and sets merchant discount m'. The per cardholder profit of the integrated bank corresponding to its cardholders' transactions is:

$$B = (1-\gamma)\left(f + a - c_I\right)\left[1 - H(f)\right] + \gamma\left(f + m' - c_I - c_A\right)\left[1 - H(f)\right],$$

where γ is A's share in the acquiring market. That is, a fraction γ of the bank's cardholders' transactions are "on us" transactions. Since the acquiring market is perfectly competitive, γ can be positive only if:

$$m' \leq m = a + c_A.$$

Then

$$B \leq \left(f + a - c_I\left[1 - H(f)\right]\right).$$

That $m' \leq a + c_A$ further implies that the bank makes no money or loses money on the transactions of cardholders of other banks who transact with the merchants it has signed up. Thus $\left(f + a - c_I\right)\left[1 - H(f)\right]$ is indeed an upper bound on the integrated issuer's profit. The issuer thus does not gain from operating in the acquiring business.[48]

48 The reader may be concerned that the conclusion follows only in the case in which the issuers are (local) monopolies. In principle, there might be strategic effects that could induce the issuing bank to raise its cost of issuing cards by losing money on the acquiring side in order to soften competition in the issuing market. It can be checked this is not so in the Bertrand and Cournot illustrations discussed just after Assumption ??. Even though the issuer loses money on its acquiring transactions, it cannot reduce this loss by losing customers on the issuing side since customers then go to another issuer and still use a card. So, even though the issuer has a higher cost, its opportunity cost of issuing cards is unaffected and there is no strategic effect. In contrast, there is a strategic effect in the Cournot case; however, this effect goes the wrong way for the integrated issuer. In the Cournot model, the integrated issuer reduces its output if it loses money per transaction on the acquiring side. But this induces other issuers to increase their own output, resulting in a further loss for the integrated issuer. We thus conclude that in either model of strategic competition, vertical integration does not increase profit.

APPENDIX B:

Observed merchant heterogeneity

To incorporate observable merchant heterogeneity, suppose indeed there are K categories of merchants (say supermarkets, grocery stores, gas stations…) parametrized by $k = 1, ..., K$. Each is characterized by (observable) merchant benefit b_S^k, and an exogenous transaction volume y^k (as earlier, we normalize total volume to 1: $\sum_k y^k = 1$). The average interchange fee, defined as

$$\bar{a} = \sum_{k=1}^{K} y^k \bar{a}^k,$$

depends only on average merchant benefit

$$\bar{b}_S = \sum_{k=1}^{K} y^k b_S^k.$$

Indeed by multiplying each equation

$$\bar{a}^k + c_A = b_S^k + \beta\left(f^*\left(c_I - \bar{a}\right)\right),$$

by y^k and summing over k we obtain:

$$\bar{a} + c_A = \bar{b}_S + \beta\left[f^*\left(c_I - \bar{a}\right)\right]. \qquad \text{(B1)}$$

Individual interchange fees $\bar{a}^k$ are then given by:

$$\bar{a}^k = \bar{a} + b_S^k - \bar{b}_S.$$

In particular, the net cost of the card is uniform across merchant categories:

$$m^{n,k} \equiv c_A + \bar{a}^k - b_S^k \equiv c_A + \bar{a} - \bar{b}_S.$$

This formula clarifies the nature of externalities between (observable) categories of merchants. Suppose for example that a new category of merchants with a low benefit (say supermarkets) participates in the system: the average benefit $\bar{b}_S$ decreases. Applying the implicit function theorem to formula (B1) shows that

$$0 < \frac{d\bar{a}}{d\bar{b}_S} < 1.$$

Therefore when supermarkets participate in the system, the average interchange fee $\bar{a}$ decreases, and the individual interchange fee of all other categories of merchants increases.

APPENDIX C:

Proof of Proposition 2

When both merchants refuse the card, the Hotelling equilibrium is symmetric and the merchants' profit is

$$\pi^* = \frac{t}{2}.$$

Suppose now that merchant *j* deviates and accepts the card. His new equilibrium profit is given by (15). Therefore the deviation is unprofitable provided that:

$$t^2 \geq \left[t + \frac{1}{3} D(f)\left(\beta(f) - m^n\right) \right]^2 - m^n \beta(f) D(f)\left(1 - D(f)\right).$$

When $a = \bar{a}, \beta(f) = m^n$ and the above relation is satisfied with a strict inequality. By continuity, it is also satisfied for a close to $\bar{a}$. Last, for $a > \bar{a}$, refusing the card is a dominant strategy for each merchant.

APPENDIX D:

Proof of Proposition 6

With cash discounts, merchants de facto compete on two segmented markets: that of consumers holding no card and that of cardholders. Let p^*_{cash} and p^*_{card} denote the two prices quoted by the merchants. These prices follow the Hotelling rule (price equals marginal cost plus the differentiation parameter):

$$p^*_{\text{cash}} = d + t$$

$$p^*_{\text{card}} = \left[d + \left(m - b_S \right) \right] + t.$$

Note that, provided that $m^n(a) = m - b_S > 0, p^*_{\text{card}} > p^* > p^*_{\text{cash}}$, where p^* is the no-surcharge price given by (5). The no-surcharge rule leads, as one would expect, to a redistribution towards cardholders.

For customer fee f, a consumer purchases a card if and only if

$$b_B \geq f + \left[p^*_{\text{card}} - p^*_{\text{cash}} \right] = f + a + c_A - b_S.$$

The key insight is that the diffusion of payment cards can no longer be influenced by the interchange fee, since the interchange fee is entirely passed through by merchants to cardholders.

To see this, let

$$\tilde{f} \equiv f + a + c_A - b_S.$$

Then the issuers' margin $\tilde{f} + b_S - c_I - c_A$ and the demand for cards $D\left(\tilde{f}\right)$ do not depend on *a*. Thus, in equilibrium, $\tilde{f}$ and market penetration, $D\left(\tilde{f}\right)$, are independent of the interchange fee.
Specifically, $\tilde{f} = f^*\left(c_I + c_A - b_S\right) > c_I + c_A - b_S$, which implies that lifting the no-surcharge rule systematically leads to an underprovision of cards.

Last, we compare payment card diffusion and social welfare under the no-surcharge rule and under cash discounts. There are more cardholders under the no-surcharge rule if and only if the net cost of a cardholder for an issuer is smaller under the no-surcharge rule:

$$c_I - a \leq c_I + c_A - b_S,$$

or

$$a \geq b_S - c_A = \underline{a}.$$

This condition is satisfied for the privately optimal interchange fee $\bar{a}$, since $\bar{a} + c_A - b_S = m^n(\bar{a}) > 0$. It is also satisfied for the socially optimal interchange fee. This is obvious in case (i) of Proposition 3 since the privately and socially optimal interchange fees then coincide. In case (ii) of Proposition 3, $f = c_I + c_A - b_S$, and so the issuers' margin, $f + \left(a - c_I\right)$, is equal to $a + c_A - b_S$. If the condition were violated, then the issuers' margin and profit would be negative, which is impossible at equilibrium. We thus conclude that card surcharges inhibit the diffusion of payment cards.

In terms of social welfare, the analysis is more complex:

- In case (i) of Proposition 3, lifting the no-surcharge rule unambiguously reduces social welfare. This is because in this case the no-surcharge rule leads to an efficient diffusion of cards, while lifting it leads to underprovision.

- In case (ii) however, the no-surcharge rule leads to overprovision $\left(f = f^*\left(c_I - \bar{a}\right) < c_I + c_A - b_S\right)$ while lifting it leads again to underprovision $\left(\tilde{f} = f^*\left(c_I + c_A - b_S\right) > c_I + c_A - b_S\right)$. Thus there is a trade-off between two sources of inefficiency: issuers' market power, which leads to underprovision of cards (when the no-surcharge rule is not imposed) and cross-subsidization between cardholders and non-cardholders, which leads merchants to refuse the card when their net cost is greater than the average (and not marginal, as efficiency would require) cardholder benefit. As a consequence, lifting the no-surcharge rule may increase social welfare when merchant resistance is weak and issuers have little market power.

REFERENCES

[1] Ausubel, L.M. "The Failure of Competition in the Credit Card Market." *American Economic Review*, Vol. 81 (1991), pp. 50–81.

[2] Baker, D.J. "Shared ATM Networks. The Antitrust Dimension." *Federal Reserve Bank of St Louis Review*, Vol. 77 (1995), pp. 5–17.

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

[4] Carlton, D.W., and A.S. Frankel "*The Antitrust Economics of Payment Card Networks*." *Antitrust Law Journal*, Vol. 63 (1995), pp. 643–668.

[5] Chakravorti, S., and W.R. Emmons "Who pays for Credit Cards?" *Federal Reserve Bank of Chicago Public Policy Series*, (2001) EPS-2001-1.

[6] Chakravorti, S., and A. Shah "A Study of the Interrelated Bilateral Transactions in Credit Card Networks." *Federal Reserve Bank of Chicago Public Policy Series*, (2001) EPS-2001-2.

[7] Chakravorti, S., and T. To "Toward a Theory of Merchant Credit Card Acceptance." Mimeo, Federal Reserve Bank of Chicago, 2000.

[8] Evans, D.S., and R.L. Schmalensee *The Economics of the Payment Card Industry*, National Economic Research Associates, 1993.

[9] "Economic Aspects of Payment Card Systems and Antitrust Policy Toward Joint Ventures." *Antitrust Law Journal*, Vol. 63 (1995), pp. 861–901.

[10] *Paying with Plastic: The Digital Revolution in Paying and Borrowing*. Cambridge, Ma: MIT Press, 1999.

[11] Frankel, A.S. "Monopoly and Competition in the Supply and Exchange of Money." *Antitrust Law Journal*, Vol. 66 (1998), pp. 313–361.

[12] Gans, J., and S. King "The Neutrality of Interchange Fees in Payment Systems." Mimeo, University of Melbourne, 2001.

[13] Gilbert, R.J. "On the Delegation of Pricing Authority in Shared ATM Networks." Mimeo, University of California Berkeley, 1990.

[14] Hausman, J., Leonard, G., and J. Tirole "The Impact of Duality on Productive Efficiency and Innovation." Mimeo, MIT, 1999.

[15] Kim, J. "The Impact of Proprietary Positions and Equity Interest in the Pricing of Network ATM Services." Mimeo, MIT, 1998.

[16] Laffont, J.J., Rey, P. and J. Tirole "Network Competition: I. Overview and Nondiscriminatory Pricing." *Rand Journal of Economics*, Vol. 29 (1998a), pp.1–37.

[17] "Network Competition: II. Price Discrimination." *Rand Journal of Economics*, Vol. 29 (1998b), pp. 38–56.

[18] Matutes, C. and J. Padilla "Shared ATM Networks and Banking Competition." *European Economic Review*, Vol. 38 (1994), pp. 1113–1138.

[19] McAndrews, J. "Retail Pricing of ATM Network Services." Working Paper n96–12. Federal Reserve Bank of Philadelphia, 1996.

[20] McAndrews, J. "The Role of Direct Presentment in Non-Cash Payments." Research Paper, Federal Reserve Bank of New York, 1997.

[21] McAndrews, J. and R. Rob "Shared Ownership and Pricing in a Network Switch." *International Journal of Industrial Organization*, Vol. 14 (1996), pp. 727–745.

[22] Priest, G. "Cartels and Patent Licence Arrangements." *Journal of Law and Economics*, Vol. 20 (1977), pp. 309–377.

[23] Rochet, J.C. and J. Tirole "Platform Competition in Two-Sided Markets." Mimeo, IDEI, University of Toulouse, 2001.

[24] Schmalensee, R. "Payment Systems and Interchange Fees." Working Paper n8256, NBER, 2001.

[25] Small, J., and J. Wright "Decentralized Interchange Fees in Open Payment Networks: An Economic Analysis." Mimeo, University of Aukland, 2000.

[26] Schwartz, M., and D. Vincent "The No-Surcharge Rule in Electronic Payments Markets: A Mitigation of Pricing Distortions." Mimeo, Georgetown University, 2000.

[27] Seade, J. "Profitable Cost Increases and the Shifting of Taxation: Equilibrium Responses of Markets in Oligopoly." Mimeo, ITAM, Mexico City, 1987.

[28] Small, J., and J. Wright "Decentralized Interchange Fees in Open Payment Networks: An Economic Analysis." Mimeo, University of Aukland, 2000.

[29] Tirole, J. *The Theory of Industrial Organization*. Cambridge: Ma: MIT Press, 1988.

[30] Vives, X. *Oligopoly Pricing: Old Ideas and New Tools*. Cambridge, Ma: MIT Press, 1999.

[31] Wright, J. "An Economic Analysis of a Card Payment Network." Mimeo, University of Auckland, 2000.

[32] "The Determinants of Optimal Interchange Fees in Payment Systems." Working Paper n 220, Department of Economics, University of Auckland, 2001.

BIBLIOGRAPHY

Ahlborn, Christian, Howard H. Chang & David S. Evans, *The Problem of Interchange Fee Analysis: Case without a Cause?* 22(8) EUR. COMPETITON L. REV. 304 (2001).

Balto, David A., *The Problem of Interchange Fees: Costs without Benefits?* 81 EUR. COMPETITON L. REV. 50 (2000).

Baxter, William F., *Bank Interchange of Transactional Paper: Legal and Economic Perspectives*, 26 J.L. & ECON. 541 (1983).

Carlton, Dennis W. & Alan S. Frankel, *The Antitrust Economics of Payment Card Networks*, 63 ANTITRUST L.J. 643 (1995).

Carlton, Dennis W. & Alan S. Frankel, *The Antitrust Economics of Credit Card Networks: Reply to Evans and Schmalensee*, 63 ANTITRUST L.J. 903 (1995).

Chang, Howard H. & David S. Evans, *The Competitive Effects of the Collective Setting of Interchange Fees*, 45(3) ANTITRUST BULL. 641 (Fall 2000).

Evans, David S. & Richard Schmalensee, *Economic Aspects of Payment Card Systems and Antitrust Policy Toward Joint Ventures*, 63 ANTITRUST L.J. 861 (1995).

Frankel, Alan S., *Monopoly and Competition in the Supply and Exchange of Money*, 66 ANTITRUST L.J. 313 (1998).

Gans, Joshua S. & Stephen P. King, *Regulating Interchange Fees in Payment Systems* (2001) (mimeo, University of Melbourne).

Gans, Joshua S. & Stephen P. King, *The Neutrality of Interchange Fees in Payment Systems* (2001) (mimeo, University of Melbourne).

Katz, Michael, *Reform of Credit Card Schemes in Australia II*, Reserve Bank of Australia (2001).

Rochet, Jean-Charles & Jean Tirole, *Cooperation Among Competitors: Some Economics of Payment Card Associations*, 33(4) RAND J. ECON. (forthcoming Winter 2002).

Rochet, Jean-Charles & Jean Tirole, *Platform Competition in Two-Sided Markets*, J. EUR. ECON. ASS'N (forthcoming 2003).

Schmalensee, Richard, *Payment Systems and Interchange Fees*, L(2) J. INDUS. ECON. 103 (2002).

Schwartz, Marius & Daniel R. Vincent, *Same Price, Cash or Credit: Vertical Control by Payment Networks* (2000) (mimeo, Georgetown University and University of Maryland).

Small, John & Julian Wright, *The Bilateral Negotiation of Interchange Fees in Payment Schemes* (2001) (mimeo, NECG and University of Auckland).

Wright, Julian, *Optimal Card Payment Systems*, EUR. ECON. REV. (forthcoming 2003).

Wright, Julian, *Pricing in Debit and Credit Card Schemes* (2002) (mimeo, University of Auckland).

Wright, Julian, *The Determinants of Optimal Interchange Fees in Payment Systems*, UNIVERSITY OF AUCKLAND DEPARTMENT OF ECONOMICS WORKING PAPER #220 (2001).

Wright, Julian, *Why Do Firms Accept Credit Cards?* (2002) (mimeo, University of Auckland).

ABOUT THE AUTHORS

Christian Ahlborn,
Partner, Linklaters & Alliance,
christian.ahlborn@linklaters.com.

The late William F. Baxter,
William Benjamin Scott and Luna M. Scott Professor of Law, Stanford University.

Howard H. Chang,
Vice President, NERA Economic Consulting,
howard.chang@nera.com.

David S. Evans,
Senior Vice President, NERA Economic Consulting,
david.evans@nera.com.

Jean-Charles Rochet,
Research Director, Institut d'Economie Industrielle, Toulouse,
rochet@cict.fr.

Richard Schmalensee,
Professor of Economics and Management and *John C Head III Dean of the Sloan School of Management*, Massachusetts Institute of Technology,
rschmal@mit.edu.

Jean Tirole,
Scientific Director, Institut d'Economie Industrielle, Toulouse,
tirole@cict.fr.

payingwithplastic.org