A CLOSER LOOK AT PAYMENT CARD SECURITY

D. Bruce Johnsen*
George Mason University School of Law
3301 North Fairfax Drive
Arlington, VA 22201-4498

703.993.8066/djohnsen@gmu.edu

Draft 1, Version 1, May 2014

* I thank Todd Zywicki for helpful discussions and the Law and Economics Center at George Mason University School of Law for generous financial assistance.

I. Introduction

This essay examines the U.S. payment card system. My primary focus is on debit cards, but I also reference close substitutes such as charge cards, credit cards, prepaid cards, and even cash and checks. I assess the security features of payment cards and their vulnerabilities. My specific goal is to determine whether and to what extent the interchange fee caps imposed under the Durbin Amendment to the Dodd-Frank Act (2011)[1] may have caused or contributed to a decline in the incentive of system participants to invest in card security, possibly increasing the system's vulnerabilities to the kind of massive data breaches recently experienced by Target and other retailers. My intuition is simple. If consumers of a good cannot assess its quality at the point of sale, a regulated price ceiling imposed on sellers is likely to lead to a reduction in quality. Security is surely a difficult-to-observe attribute of payment processing and is therefore a viable candidate for underinvestment by system participants in the face of a binding cap on interchange fees.

The economic literature on payment systems began with William Baxter's seminal *Bank Interchange of Transactional Paper: Legal and Economic Perspectives* (1983). The paper proposes that payment systems constitute "two-sided markets." Others, including Evans and Schmalensee (1995) and Rochet and Tirole (2002), have since formalized the analytics behind two-sides markets. The essential attribute of two-sided markets is that they must coordinate the demands of two different groups of consumers. One-sided markets coordinate across buyers and sellers of real goods, most fundamentally in a barter system. But in any system more complex than barter it is widely acknowledged that both buyer and seller are also consumers of a transaction facilitated by an intermediary, whether it be mere pairing of buyers and sellers or the processing and underwriting of payments. Newspapers, by way of example, sell both content to readers and ad space to advertisers. Presumably the two are complements. But what is the optimal price of content to readers and what is the optimal price to advertisers for ad space? The only constraint under competition is that the total price to

[1] *Dodd–Frank Wall Street Reform and Consumer Protection Act* (Pub.L. 111–203, H.R. 4173).

both parties just cover the cost of the transaction. Without readers, advertisers will be unwilling to pay for ad space, while without ads newspaper readers might be close to indifferent. Efficient pairing requires that readers pay a low price for the newspaper and that advertisers bear most of the publication costs in the fees they pay to advertise.

Some have suggested that readers end up paying less than the marginal cost to the publisher of providing content, while advertisers pay more than the marginal cost of ad space, with advertisers in essence cross-subsidizing readers. The literature recognizes, however, that the good in question—in this case the pairing of readers with advertisers—is a joint product. As with cattle, which generate both meat and hides, there is no economically correct way to allocate production costs between meat and hides, and there is no reason to expect some kind of superficial parity between the price of meat and hides.

In the context of payment systems, the buyer/cardholder and the seller/merchant each consume, or receive the benefit of, the transaction, including payment processing and underwriting. Nominally, the merchant pays the transaction cost in the form of what is called the merchant discount, equal to the price of the good minus the cost of processing the payment. If the cardholder pays $100, the merchant's account may be credited with about $98, with much of the difference going to the bank that issues the card to the holder, underwrites the payment process, and monitors security. It appears as though the cardholder pays very little of the two-dollar transaction cost and may even be subsidized to use the card by way of cash or in-kind rewards, unpriced float, etc. The literature shows that this is necessary to equalize the elasticity of demand for transactions on the part of cardholders with the elasticity of demand on the part of merchants. Merchants want buyers, and card availability and acceptance help get them. Buyers want merchants that accept the card or cards they carry. The optimal solution is for merchants to honor all cards within a payment system to attract buyers. Coordinating everyone to "get on board" is the function of the specialized card association such as Visa or American Express.

Whether it is inevitably true or not, two-sided markets appear to exhibit indirect network effects. The more merchants that accept the card they carry, the better off cardholders are. The more cardholders there are who carry a card the merchant accepts

the better off the merchants is. The solution that minimizes transaction costs appears to be one in which the card association coordinates the payment system to clear the market, setting the terms of trade to the merchant and allocating the merchant discount between various parties involved in the processing, underwriting, and security functions.

I take no issue with the analytics of two-sided markets. But at the same time I am somewhat suspicious about whether they represent a fundamental departure from so-called one-sided markets in which transacting is costly. Quite possibly, the only departure is that in one-sided markets the relevant functions are bundled into, or vertically integrated with, the trade of real goods and manifested in the difference between the nominal price the parties pay and the net value, after transaction costs, they each capture. What we call two-sided markets may simply reflect specialization in bearing the potential variability in transaction values across multiple dimensions, one of which, of course, is payment card security.

I begin in Part II by describing the payment card system in greater detail and showing how it fits with the literature on two-sided markets. My approach, here, is to go back to basics by bringing to bear some of the early and subsequent literature on transaction costs. Part III presents a brief literature review. Coase's (1937, 1960) seminal work on the firm and the cost of transacting provide the foundation, but contributions by Demsetz (1968) on the demand for immediacy in transacting, Hishleifer (1971) on excess search, Barzel (1976) on taxes, Alchian (1977) on money, and Klein and Leffler (1981) on quality assurance, all provide useful insights into the payment card system. My hope is that this review will identify the relevant tradeoffs at stake from the imposition and regulation of interchange fee caps. Part IV explains the fee cap and other restrictions the Durbin Amendment and subsequent Federal Reserve rules imposed on debit card transactions. I then assess their likely effect on security and consumer welfare throughout the payment card system. Part V provides a brief summary and concluding remarks.

II. The Payment Card System and Two-Sided Markets

A. Payment Card System Overview

The baseline for understanding payment systems is barter in one-sided spot markets. Buyers and sellers of real goods negotiate terms of trade and no doubt incur substantial transaction costs assessing or guaranteeing the quality of the goods and policing performance. Any payment system, whether cash, check, or plastic, must involve lower transaction costs than barter or the parties would refuse to participate in it.

Baxter (1983) traces the evolution of transactional paper in the U.S. from the time of the early national banking system through development of credit cards. Early forms of transactional paper consisted primarily of bank notes and bank drafts. For the buyer of a good or service from outside a given locality to arrange payment to a distant seller in anything but cash required a byzantine series of transactions involving the buyer's and seller's banks and often one or two others, with each bank charging a fee for "exchange" along the way. Legal tender at the time of the new republic was scarce. As cumbersome as the transactional system was, at the margin it must have been better than barter or cash, and in any event it gradually evolved to become fairly streamlined. With prodding from the newly created Federal Reserve Board starting in 1913, exchanging banks increasingly began to accept one another's bank drafts—personal checks—at par. This is because whatever risks of nonpayment and other adverse events they bore largely balanced out across banks.

The use of plastic started in the 1950s with the rise of the first three-party or closed-loop charge card systems pioneered by American Express and Diners Club. These banks found it profitable to issue cards to affluent depositors, who demanded liquidity for business travel. The banks would sign up, or "acquire," likely issuers such as prominent hotels and restaurants to honor the cards, with a guarantee that the bank as underwriter would pay the cardholder's debt. The cardholder was required to pay the entire balance monthly and enjoyed the float in the meantime, apparently as an inducement to carry the issuer's card. Merchants were willing to accept a discount from retail receipts and delayed payment in exchange for the additional sales to cardholders.

The predecessor of the MasterCard and Visa credit card associations appeared in the 1960s. These associations began as mutuals owned by various regional member banks to coordinate the exchange system by setting and policing fees and other terms of

exchange. Eventually, MasterCard and Visa went public. The system in which they operate is characterized as four-party, or open-loop because it vertically disintegrates the coordination function from the issuing, underwriting, acquisition, security, and processing functions. Figure 1 illustrates the approximate structure of the open-loop payment card system. A transaction begins when the cardholder buys goods or services from the merchant and presents an association-branded card the merchant accepts for payment. The card has been issued to the holder by the issuing bank in cooperation with the branded association. The system performs two basic functions, authorization and clearing and settlement. What is called a four-party system is really at least a five party system including the cardholder, the merchant, the acquiring bank, the issuing bank, and the branded association.

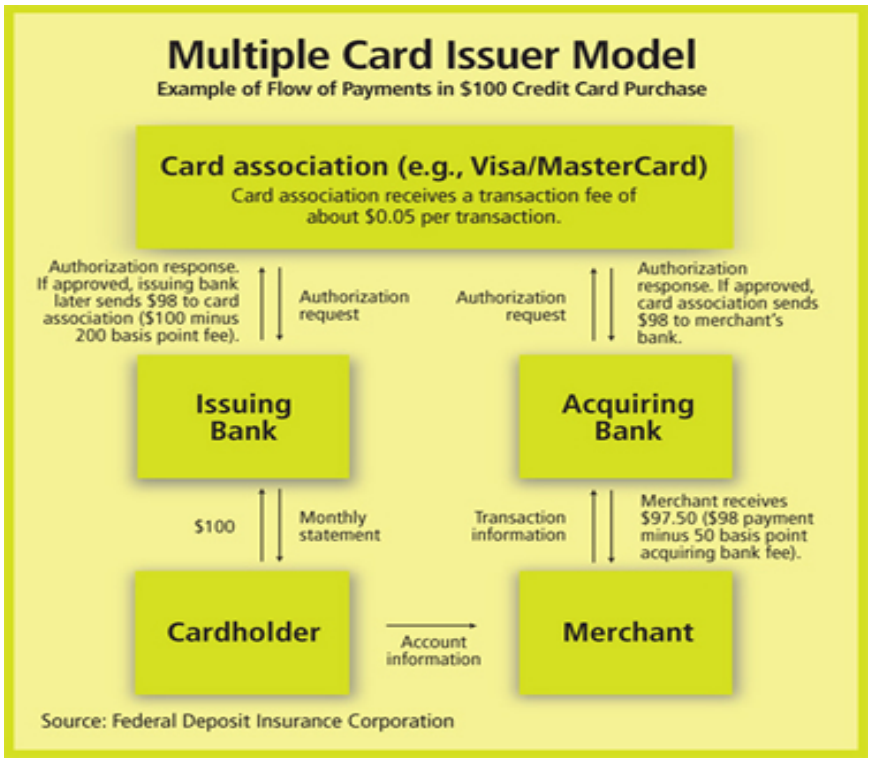

Figure 1: *The Payment Card System*

In fact, the open-loop system is considerably more complex than what Figure 1 suggests. Any number of specialized intermediaries, most importantly electronic aggregators and technical network administrators have arisen to facilitate payment

processing both on the front end and the back end. It is unclear how much these intermediaries are paid or how the payments they receive are metered, but the process can be broken down into the authorization, aggregation, and clearing and settling functions as described below.[2]

Authorization: The cardholder presents the association-branded card containing his account information to the merchant at the point of sale. The merchant or cardholder swipes the card into a point-of-sale electronic terminal or a payment gateway through a secure connection from a website, retail location, or a wireless device.[3] The payment gateway, which may be an independent firm as well as part of an electronic network, receives the secure transaction information and passes it through a secure connection to the acquiring bank's front-end processor. Also apparently a separate firm, the front-end processor aggregates authorization requests and submits the transactions to the card association network, which routes them to the cardholder's issuing bank. The issuing bank approves or declines the transactions and passes the results back through the association network, which relays the results to the acquirer's front-end-processor and through the point-of-sale terminal or payment gateway. The payment gateway stores the transaction receipts and sends them to the merchant, which receives the authorization response and completes the transaction accordingly.

Clearing and settlement: The merchant deposits the transaction receipts with the acquirer by way of a settlement batch. The captured authorizations are passed from the front-end network to the back-end network for settlement. The back-end-processor generates ACH (automated clearing house) files for merchant settlement and sends them to the acquiring bank, which credits the merchant's line-of-credit account, normally by the end of the business day. The acquiring bank submits settlement files to the issuing banks for reimbursement via the interchange network. The issuers post the transactions to the cardholder accounts and sends cardholders a monthly statement.

B. The Economics of Two-sided Markets

[2] See, e.g., https://www.ippay.com/index.php?q=merchant_processing_overview.
[3] It appears that a point-of-sale terminals are used exclusively by brick-and-mortar merchants, while payment gateways may be used by both brick-and-mortar and online merchants.

- Review Baxter, Rochet and Tirole, Zywicki
- Network hold-up by issuers
- Benefit from default interchange fees set by the card association
- Relative demand elasticities by merchants and cardholders

III. The Transaction Cost Literature

The total cost of payment card exchange is a transaction cost, albeit one subject to considerable specialization and economizing. To evaluate the likely effect of regulated ceilings on debit card interchange fees it is worth reviewing the literature on transaction costs. My review starts at the beginning but proceeds in a logically ordered fashion rather than in chronological order of publication.

The literature began as we know it with Coase's (1937) early work on the theory of the firm. According to Coase, using prices to intermediate market exchange is costly. We can think of these costs as transaction costs. The firm arises to economize on the costly use of prices. By way of example, market exchange requires the parties to negotiate terms of trade, to meter what they expect to give up and to get, and to enforce the terms of trade, all of which incur transaction costs. What is more, by metering exchange through revealed prices one or both of the parties might reveal entrepreneurial secrets that reduce their ability to capture returns on invested capital (Johnsen, 2001). On one hand, by organizing activity within the firm rather than in the market some transaction costs can be avoided. On the other hand, exchange within the firm under the direction of the entrepreneur is also costly, even though it avoids metering, and these costs can also be considered transaction costs. The extent of the firm is determined by balancing the transaction costs of internal unpriced exchange with the transaction costs of external, or priced, market exchange.

The difference between closed-loop and open-loop payment card systems clearly reflects a difference in the extent of the firm. American Express integrates system coordination, payment processing, card issuance, card security, and merchant acquisition.

The only revealed price is the merchant discount. How these different functions get performed within the firm and what remuneration they command is all determined inside a black box. In an open-loop system the functions are more transparent and are, to a significant extent, separately priced. In gross, the merchant pays a discount for exchange processing, and this discount is shared between various separate firms. Visa acts as the central contracting agent (Alchian and Demetz, 1972), branding cards, coordinating fee allocations, penalizing participants, paying incentives, and establishing security standards. The acquiring banks aggregate and underwrite transactions and provide security, and the issuing banks underwrite cards and provide security. In fine, other parties participate in the system as front-end and back-end network aggregators and processers and receive a slice of the merchant discount as well. Perhaps what is "open" about the open-loop system is that the participating firms have no exclusive role, but instead face competition from other firms. It seems truly peculiar that the more transparent open-loop system should suffer fee regulation for a function that is completely opaque in the closed loop system. Perhaps the lesson is that one cost of transacting in the market is the risk of having your prices regulated. Long live vertical integration!

Coase's (1960) path breaking work on social cost changed the way economists look at market exchange, contract terms, economic organization, and the contours of legal rules. The so-called "Coase Theorem" says that if transaction costs are zero the structure of property rights is irrelevant to the allocation of resources. Economic efficiency rather than initial ownership positions determine resource allocation. Coase never used or promoted the term Coase Theorem,[4] and nor did he consider a world of zero transaction costs a relevant benchmark for economic analysis. Rather, his point was that to understand why the structure of property rights matters to resource allocation we must focus on the cost of transacting as an explanatory variable. One important implication is that any inefficiency economists might imagine on a blackboard is a profit opportunity for market participants. Markets and other economic institutions exist to capture these opportunities, but the cost of transacting gets in the way. If a blackboard

[4] The term Coase Theorem is attributed to George Stigler.

inefficiency persists in the real world it must be because the transaction costs the parties face to eliminate it exceed the benefits from doing so. Transaction costs are real costs to be avoided just like production costs.

By way of example, so-called "information asymmetry," in which two parties engage in trade even though one party is a specialist and therefore better informed than the other, is often bemoaned as an example of market failure and social inefficiency. But it is quite the contrary in Coasean terms. If parties routinely transact under conditions of information asymmetry it must be because the form of organization within which they trade allows the uninformed party to trust the informed party. The uninformed party can therefore economize on the cost of becoming informed and joint surplus rises. Why waste resources by duplicating information? To observe a persistent information asymmetry is to observe a successful market at work.

On its face, the price ceiling and other regulations the Federal Reserve imposed on open-loop debit card interchange fees (the fee allocation to the issuing bank) under the Durbin Amendment seem inexplicable in Coasean terms. Although not zero, transaction costs within the payment system must be low compared to the alternatives, such as cash or checks.[5] The very reason the payment system exists is to reduce the costs of exchange. And the system has shown itself to be innovative and entrepreneurial, with declining transaction costs over time.[6] For the Durbin Amendment to enhance either cardholder or merchant welfare, it must have been true that the parties were leaving money on the table. Doing so is not part of their business model. Others have noted the apparent absence of any kind of market failure that might justify regulation (Rochet and Tirole, 2002).

Alchian's excellent exploration *Why Money?* (1978) provides important insights into payment systems, one example of which is money itself. The article shows that, unlike traditional lore, money is neither necessary nor sufficient to solve the problem of "double coincidence of wants" but only to solve the problem of "double coincidence of information." He images a simplified economy consisting of novices and experts in four goods, diamonds, oil, wheat, and *C*. All parties know the value of their own goods.

[5] We would expect the cost of various payment systems currently in use to be equal at the margin. Most consumers divide their payments between cash, check, and plastic, and most merchants divide their receipts similarly.
[6] Cite Zywicki?

Absent money, novices must jointly incur substantial transaction costs to assess the quality of any good they might barter with another novice, and vice-versa, which substantially limits the joint surplus from trade. Bartering with an expert in any good involves lower transaction costs under the assumption that the expert can assess the value of the good for which he is an expert at low cost and that he can be trusted to convey this value accurately to buyers and sellers. If a diamond novice seeks to trade his diamonds (whose quality he knows) for a wheat expert's wheat, the joint transaction costs are substantially lower than if the diamond novice trades with the wheat novice.

Even if there is a good *C*, cash, for which all parties have relatively low assessment costs, there is no benefit to the novice from interposing the trade of diamonds for cash then cash for wheat, whether with a novice or an expert, since all this does is incur another layer of transaction costs. The diamond novice is better off going directly to the wheat expert to satisfy his wants. The costs of transacting can be further reduced, however, if the diamond novice sells his diamonds to the diamond expert in exchange for cash (which the diamond expert gets at low cost from the cash expert) and then takes the cash to the wheat expert to buy wheat. In Alchian's words, the informational gains are "the result of [the] ability to get quality assurance at a lower cost from the diamond and wheat experts without imposing on them the higher costs of identifying goods other than *C*, in which most people are nearly experts." Having access to a universally low-assessment-cost good, money, together with the ability to trade real goods with multiple specialists at low cost increases the social trading surplus and encourages people to specialize.

Aside from the obvious lesson that there are benefits from specializing in transactional exchange in two-sided markets, there is another important lesson from Alchain's simplified economy. When novices go to barter real goods, they are also mutual demanders of "the transaction," just as in two-sided markets, and they jointly incur the cost of transacting. With novices in n goods whose assessment costs vary across the remaining n-1 goods, a diamond novice who wants wheat but faces high costs assessing wheat may end up trading diamonds for oil and leaving it at that, even though in some abstract sense he prefers to have wheat. He balances not just the relative valuations he places on the two goods but the relative assessments costs that are jointly

incurred with the contra party. This calculus also applies to the contra party. Barter markets therefore strike me as two-sided markets that perform the same balancing act that happens in any other two-sided market, with the difference being that there are no transactional specialists in the barter market and no explicit pricing of the transaction, itself, for us to observe. The balancing function is vertically integrated into to what appears to be the simple exchange of two real goods, and this exchange generates value (and economizes on costs) on multiple dimensions.

This is not a criticism of the literature on two-sided markets, but it does suggest that in assessing any payment system it helps to take notice of the broader transaction cost literature. Among other things, this literature has unpackaged the multi-dimensionality of trade in real goods. Barzel's (1976) work on taxation provides a compelling example. It describes the effect on the incentives of consumers and producers in responding to a tax, not as participants in a one-sided market, but as joint consumers of a transaction in a two-sided market. With the imposition of a per unit tax, say 20 cents per pack on cigarettes, the parties to the transaction have a mutual interest in cooperating to reduce the incidence of the tax. The legal definition of the good being taxed, "packs," does not fully encompass the multi-dimensional nature of the value and cost involved in transacting the underlying economic good. Pre-tax, consumers and producers cooperate to maximize the gains from trade by optimally balancing values and costs on multiple dimensions. During the early 1970s, this lead to trade in standard packages of 20 cigarettes, each 88 millimeters long, of given quality, with in-time delivery to the merchant in packaging suitable to preserve freshness, etc., etc., etc.

With imposition of the tax, consumers and producers had an incentive to put more tobacco—or smoking pleasure—into every pack. Although this increased the marginal value to consumers, the marginal cost to producers, and price of a pack of cigarettes, it allowed the parties to transact a given amount of smoking pleasure by trading fewer packs of 100 millimeter cigarettes, thereby reducing the total tax burden.[7] The new characteristics of the pack, which were sub-optimal pre-tax, are now locally optimal. This is exactly the kind of balancing the model of two-sided markets is designed to solve.

[7] Note that, with these adjustments possible, the price of the new pack might increase by more than the amount of the tax, a result inconsistent with standard textbook analysis of one-dimensional goods.

A look at *ad velorem* taxes is also revealing. If the tax on cigarettes had been five per cent of the purchase price rather than 20 cents per pack, the adjustments described above would have hurt rather than helped the parties. This is because the price of the new pack was higher than the price of the old pack, and with *ad* velorem taxes this would have increased the total dollar tax per pack, negating the possibility of tax savings. What adjustments could the parties make to an *ad velorem* tax? As before, any good produces value and incurs costs on multiple dimensions. Just as it is possible to bundle valuable attributes into a good, it is possible to unbundle valuable attributes and transact them separately in an untaxed setting. By way of example, with an *ad velorem* tax on cigarettes the parties could switch to less costly, less valued, and lower priced unfiltered cigarettes, leaving consumers to buy the filter in a separate untaxed transaction. Perhaps a more salient example is the effect of the *ad velorem* TSA tax on airfares following 9/11. The tax applies only to the ticket price. It is therefore unsurprising that with imposition of the TSA tax the airlines immediately began unbundling the bundled-in attributes of the pre-9/11 good by charging separately for baggage and reducing airfares to reduce the tax burden. Doing so was in the joint interest of both passengers and air carriers.[8]

Barzel's analysis of taxes further emphasizes the point that so-called one-sided markets must be treated as two-sided markets as long as transacting is costly and the objective of the analysis is to identify the effect of transaction costs on the parties' choices. Most important, the transacting parties are not passive participants in the exchange. At any point in the exchange that meters the transfer of value between parties, there will be opportunities for value capture.[9] The tax example provides a simple though compelling demonstration, perhaps because the working of government taxing authorities is often unparsimonious, at least at inception. If anything good can be said of regulation it is that regulators generally have only a primitive understanding of the markets they regulate, leaving market participants with ample room for adjustment and economists with another project for study.

[8] While it is true that airfares did not fall, it is sufficient that they increased less than otherwise.
[9] Passengers might overload the carry-on capacity of the aircraft and carriers might cancel poorly subscribed flights under the guise of mechanical failure.

Private wealth maximizing parties recognize the ambitions of their counterparts and will seek to pre-empt any opportunities for wealth capture if doing so generates benefits in excess of transaction costs. This is exactly what allows economists to explain the form of organization the parties choose. With respect to the card payment system, it is necessary to identify how value is metered in each sub-transaction and what possible dimensions of adjustment for capturing value the parties have at their disposal along the way. In Part IV I discuss how the Durbin Amendment's regulatory cap on interchange fees might have affected payment system security and cardholder welfare. But I also discuss the likely effects of mandating per unit interchange fees where *ad velorem* pricing was the primary method of metering pre-Durbin.

One last note on taxes; it is well established in the economics of taxation that who—whether buyer or seller—formally pays the tax is irrelevant to who bears the burden of the tax. Tax burden is determined by the relative elasticities of demand and supply, with buyers bearing less of the burden as the price elasticity of demand increases, and vice-versa. Critics of debit card interchange fees, including merchant groups, note that merchants "pay" the total transaction costs in the form of the merchant discount. But whether, or two what extent, they bear the entire burden is an open empirical question. The answer depends in part on the effect of accepting a given card on merchant sales. It would be peculiar to suggest that merchants bear the entire two percent merchant discount if the net margin on incremental sales exceeds two percent. As with any activity, one cannot expect to enjoy the benefits without bearing the costs. Presumably, merchants' use and acceptance of debit cards provide all parties with benefits while requiring all parties to bear costs, increasing net benefits to all. If not, any party is free to revert to cash, or credit, or, following in the wake of Durbin, prepaid cards. In this sense, the analogy to tax incidence is misleading because the costs incurred by each party along the way are endogenous to the benefits they receive. Having to pay a price, however metered, is better than not getting the good at all. In this sense the tax analysis is inapt.

One of two considerations Alchian assumes away is the possibility that transacting parties will spend too much assessing the quality of the goods they trade in an attempt to capture wealth from contra parties. Hirschleifer (1971) shows that excess search by trading parties can reduce the joint surplus from exchange. Imagine a grocer

who sells produce of given average quality and hosts a regular clientele of shoppers. Each morning the grocer puts out a bin of apples of average quality and prices them accordingly. Shoppers filter in and sort among the apples in the bin because some apples are better than average and others are worse than average. Assuming shoppers have skill in sorting, and assuming for simplicity that they have identical relative valuations for various apple characteristics, the early shopper will buy a sample of better-than-average quality, leaving the remaining apples of lower quality than average. As the day progresses, the grocer must either lower the price to reflect the remaining average quality or suffer apple spoilage. Early shoppers gain at the expense of late shoppers between each price-adjustment. Shoppers are likely to rush to the grocery store. Not only does rushing and sorting by shoppers incur real costs, but price adjustments by the grocer incur real costs.

Sorting of this kind is costly but generates no social value. It is the equivalent of spending resources to transfer wealth from one party to another. The costs of sorting feed in to reducing the grocer's profit and, in the long run, increase the average price he charges for apples. His preference would be to have customers commit to buying blind rather than picking and choosing. If this could be accomplished, both the grocer and his customers could be made better off. Picking-and-choosing by customers is a cost of transacting to be avoided if possible, and this motivation allows economist to explain the parties' choice of contract terms (Barzel, 1977).

Applying this to Alchian's simplified economy, novices might spend too much to evaluate quality when transacting with another novice. When two novices trade, both are likely to incur assessment costs in an attempt to capture wealth from the other. The avoidance of duplicate search is most likely the source of gains from transacting with experts. What experts provide is not perfect certainty as to the value of the good for which they are experts, but increased accuracy. Like all goods, accuracy in valuation has both costs and benefits. Optimal accuracy requires traders to equate the marginal cost of accuracy with the marginal benefit. But in the grocery store example customers face a dissipating open-access race to first possession. There are various contract terms the parties might adopt to minimize the resulting dissipation. The grocer, for example, might bundle apples together randomly in transparent bags, thereby limiting customers' ability

to pick and choose and circumventing their pre-purchase inspection. It might also provide a limited warranty that reduces consumers' incentive to inspect by promising ex post adjustments in the terms of trade. Finally, it might limit customers' ability to profit from picking-and-choosing by limiting the quantity of the good they can buy. Although it seems implausible, the first buyer of apples in the grocery store could engage in arbitrage buy buying the best apples and then selling them outside the grocery store at a premium price. Lest this example be fanciful, consider the enmity concert promoters have for ticket scalping.

Picking and choosing can be a real problem in any card payment system. Absent constraining regulation, the discount the merchant pays on various cardholder transactions differs, even within a given card brand. It would be tempting for merchants to spend resources to select in favor of low-discount cards or to impose up-charges on high-discount cards. At each step in the payment card network, participants likely have opportunities to engage in socially wasteful search, and it should be in participants' joint interest to avoid this dissipating activity. It is largely the function of the branded card association to set mandatory and default rules to this end.

A second consideration Alchian assumes away is the problem of quality assurance. In his simple economy, experts are assumed to provide quality assurance at zero identified cost. This is good enough for the points he wants to make, but we now have compelling models that show how market participants address the quality assurance problem. Klein and Leffler (1981) develop a model in which consumers buy goods whose quality they cannot perfectly assess at the point of sale. Such goods are known in the economics literature as experience goods. The problem with experience goods is that the seller might promise to provide high-quality units of the good at a price sufficient to cover his costs, while secretly cutting quality to reduce costs. To the extent consumers can be fooled in this way the seller can make a one-time profit until they catch on, leaving them worse off than if they had refused to pay for high quality from the start.

This situation is illustrated in Figure 2. For simplicity, assume unit sales are strictly a function of time, shown on the horizontal axis. Value, cost, and price are shown in dollars per unit on the vertical axis. At Time 0 consumers begin paying the seller $P_H = MC_H$ for the high quality good. The seller earns no profit, just covering his opportunity

cost for the high-quality good. At Time 1 the producer cheats by lowering cost to MC_L. If it takes until Time 2 for consumers to catch on and terminate sales, the producer can earn a one-time profit equal to the double cross-hatched box. Thereafter, consumers will refuse to pay any price above P_L. Anticipating this outcome, they might refuse from the start to pay any price in excess of the cost of the low-quality good. Trade in the high-quality good would never occur, even though it could increase the social surplus. A so-called "lemons" market prevails (Akerloff, 1970).

The solution is for consumers to offer, and the seller to accept, a premium price for the high-quality good in excess of its production cost. This price is illustrated by P* in Figure 2. A seller who receives P* per unit and incurs costs equal to MC_H per unit can also cheat. If he does so, he captures a one-time profit equal the single cross-hatched box. Although the gain from cheating is higher than before, he now faces the prospect of losing the flow of surplus income equal to $P^* - MC_H$ from Time 2 to Time ∞. This flow is a perpetuity whose present value at Time 1, the moment at which the producer considers cheating, is $[P^* - MC_H]/r(1+r)$, where r is the appropriate discount rate. For given discount rate and given delay before consumers detect cheating, there is some premium price sufficiently high that the lost perpetuity exceeds the producer's one-time gain from cheating. Premium product prices therefore assure consumers they will get the high-quality good. As in any consumer advocacy setting, mandating lower interchange fees on debit transactions does not necessarily help consumers if the good in question is an experience good.[10]

This does not quite end the story because it may leave the producer with a surplus, which cannot persist in competitive equilibrium. Competing producers will vie for customer business but cannot do so by cutting price. Price cuts signal low quality. Instead, producers will compete by investing any surplus in capital to signal the quality of their product. The capital must be entirely sunk in the sense that it can have no value in the market if the producer is caught cheating. Subject to this constraint, it must provide the highest possible value to consumers. The obvious example in the context of card payment systems is the brand of the card association, e.g., Visa. The brand, costly to

[10] Recall that merchants and cardholders are both consumers of the transaction.

establish and worthless if the card association cheats, tells consumers they will receive the difficult-to-assess but valuable attributes they expect, including card security.

To better understand the card payment system, it helps to look at work from transaction cost economics on other two-sided markets. Demsetz (1968) provides an insightful analysis of transaction costs on the New York Stock Exchange (NYSE). Buyers and sellers of securities on the NYSE have a demand for "immediacy" in their transactions. They may be anxious to clear a trade, among other reasons for fear that their private information will leak out and cause adverse changes in the price of the security before they can get the trade done. But how do they know at any give moment there will be a contra party in the market willing to take the other side of the trade at the same price and in the same volume?

The NYSE specialist in ABC stock performs this function. He receives standing limit orders ("buy X shares of ABC stock at any price up to but no greater than $60," or "sell Y shares of ABC stock at any price no less than $60.25") from floor brokers on both sides of the market. Based on his knowledge of aggregated limit orders he posts the price at which he is willing to buy, his bid price, and the price at which he is willing to sell, his ask price. The bid-ask spread is the difference between his ask price and his bid price. As the day progresses he adjusts the bid-ask spread according to the arriving limit orders and either facilitates the trades of floor brokers or trades for his own account. Whenever he trades for his own account simultaneously he earns the bid-ask spread. He might buy without simultaneously selling, or sell without simultaneously buying, in which case he bears the risk of adverse price moves on his inventory. If he is any good at what he does, however, he typically makes money because of his superior private information. He fulfills his duty to make an orderly market by matching buyers' and sellers' demands and in the process provides them with the valuable immediacy they desire.

It is not too hard to see the parallel to the payment card system. Merchants and their customers both demand transactional immediacy. To transact using cash, the buyer must first get the cash from his bank, then take it to the merchant and tender it. The merchant must count it at the point of sale and provide the customer with a receipt to evidence title and as a record to facilitate any returns consistent with the merchant's return policy. At the end of the day, the merchant must count his entire till, fill out a

deposit slip, and take both to its bank, probably no sooner than the next business day. At the bank, the deposit must again be counted. Considerable time will pass from the moment the customer makes his buying decision until the purchase price finds its way into the merchant's bank account. With payment cards in electronic networks, all this happens in a New York minute. Authorization is almost instantaneous. And the merchant will have his account credited by the end of the day without having to bear the costs of float or the buyers's credit risk.

Both the cardholder and the merchant benefit from the immediacy payment cards provide, just as securities buyers and sellers benefit from the availability of specialists on the NYSE. The merchant discount is simply the bid-ask spread and compensates the intermediaries in the system, in part, for supplying immediacy. Like the specialist, they bear credit risk, the risk of fraud, float costs, etc., that would otherwise take time to resolve before the transaction could be cleared. In neither market is it possible to determine who bears what share of the transaction cost burden at any given moment. A priori, it is no more correct to say that merchants bear the burden because they suffer the discount than it is to say that cardholders bear the burden because they pay a mark-up for the goods.

- Add discussion of Barzel on measurement costs, Cheung on price ceilings, and Klein, Crawford, and Alchian on opportunism.

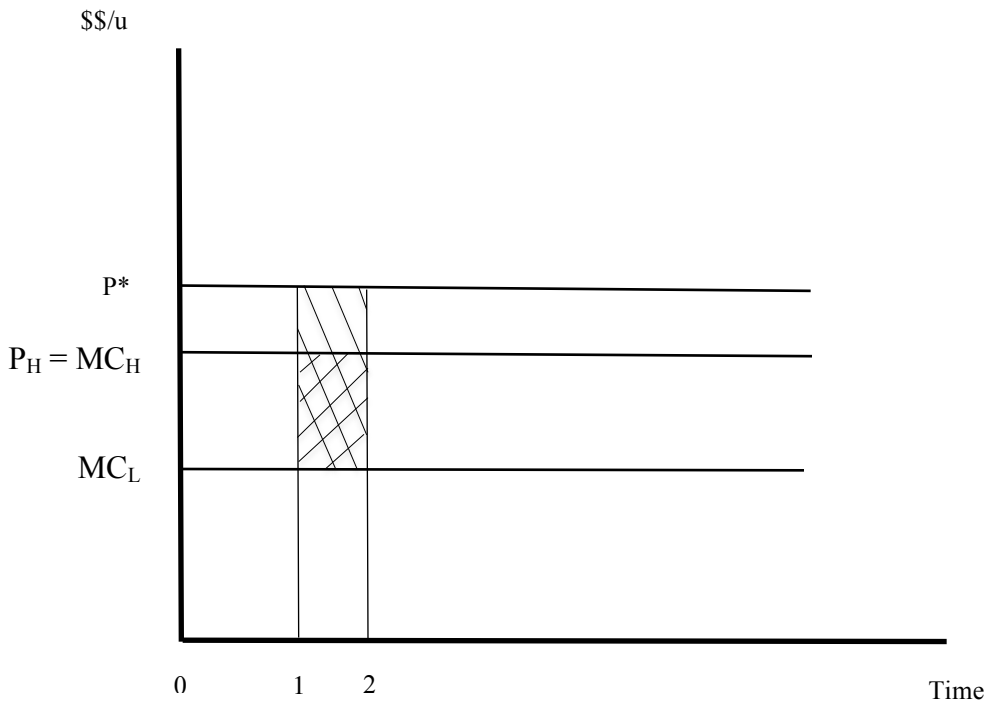

Figure 2: *Quality Assurance*

IV. The Durbin Amendment and its Likely Effects

- What did Durbin do?
- Return to the per unit, ad velorem analysis. The cap on debit interchange fees for large banks not only reduced interchange fees but it dramatically changed they way they are metered. Both have predictable effects.
- Note the effect on quality, and quality assurance, from a price cap on what is surely an experience good.
- Note that smaller banks are unlikely to be the drivers of card security innovation. By capping the fees large banks can earn, card security is likely to advance more slowly.
- Likely effects include limitations on card issuance, substitution into credit cards, pre-paid cards, cash, checks, etc. What effect on fraud, but also on overall consumer welfare. Fraud prevention is not the only thing consumers value. What about lost convenience?
- Another likely effect is that issuers will summarily decline a greater share of authorizations. The cardholder wants more security holding all else constant. But one way for banks to respond to fee caps is to unbundle parsimonious authorization. If it is likely I will be declined on my trip to Italy, I, as the cardholder, may bear the cost of picking up the phone to notify my issuer that I will be traveling there. What effect did Durbin actually have on the rate of declined debit card transactions?

V. Summary and Concluding Remarks

- Careful analysis of the payment card system using transaction cost economic should be revealing once the exact details of the system are known. Most important is to understand who captures what components of the merchant discount, how those payments are metered, and what unpriced margins the various parties have on which to make maximizing adjustment.

LIST OF REFERENCES

Akerlof, George A., *The Market for "Lemons": Quality Uncertainty and the Market Mechanism*, 84 Q.J.E. 488 (1970).

Alchian, Armen A., *Why Money?*, 9 J. MONEY, CREDIT AND BANKING 133 (1977).

Alchian, Armen A. and Allen, Exchange and Production: Theory in Use

Alchain, Armen A., and Harold Demetz, *Production, Information Costs, and Economic Organization*, 62 A.E.R 777 (1972).

Barzel, Yoram, *An Alternative Approach to the Analysis of Taxation*, 84 J.P.E. 1177 (1976).

Barzel, Yoram, *Measurement Cost and the Organization of Markets*, 25 J. LAW & ECON. 27 (1982).

Barzel, Yoram, Habib, Michel A., and D. Bruce Johnsen, *Prevention is Better Than Cure: The Role of IPO Syndicates in Precluding Information Acquisition* 79 J. BUS. 2911 (2006).

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

Coase, Ronald, H., *The Nature of the Firm*, 4 ECONOMICA 386 (1937).

Coase, Ronald H., *The Problem of Social Cost*, 3 J. LAW & ECON. 1 (1960).

Demsetz, Harold, *The Cost of Transacting*, 82 Q.J.E. 33 (1968).

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

Hirshleifer, Jack, *The Private and Social Value of Information and the Reward to Inventive Activity*, 61 A.E.R. 561 (1971).

Johnsen, D. Bruce, *Customary Law, Scientific Knowledge, and Fisheries Management among Northwest Coast Tribes*, 10 N.Y.U. ENVIRONMENTAL LAW JOURNAL 1 (2001).

Klein, Benjamin, and Keith B. Leffler, *The Role of Market Forces in Assuring Contractual Performance*, 89 J.P.E. 615 (1981).

Rochet, Jean-Charles and Jean Tirole, *Cooperation among Competitors: Some Economics of Payment Card Associations*, 33 RAND J. ECON. 549 (2002).