A CLOSER LOOK AT PAYMENT CARDS

D. Bruce Johnsen,
George Mason University School of Law

George Mason University Law and Economics Research Paper Series

16-06

A CLOSER LOOK AT PAYMENT CARDS

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 3, February 2016

* For helpful comments I thank Yoram Barzel, Todd Zywick, and participants in the Robert A. Levy Fellows Workshop in Law & Liberty at George Mason University School. The Law & Economics Center's Data Privacy and Security Project at George Mason University School of Law provided financial assistance.

A CLOSER LOOK AT PAYMENT CARDS

D. Bruce Johnsen

Abstract

This essay takes a closer look at the U.S. payment card system, primarily debit cards. I examine the bundle of transactional services this and other types of payment cards provide. My goal, in large part, is to assess the competitive effects of the debit card interchange fee cap under the Durbin Amendment to the Dodd-Frank Act (2011). In addition to a binding fee cap, it mandated a change in the way the fee is metered. A maximum per transaction fee of 20 cents, binding for most transactions, replaced a typical two-percent negotiated fee. I test hypothesis that the cap caused or contributed to a decline in the willingness of payment card intermediaries to invest in security, possibly increasing the system's vulnerabilities to the kind of data breaches that have become ever more commonplace.

I. Introduction

This essay takes a closer look at the U.S. payment card system, a so-called two-sided market. My primary focus is on debit cards, but for comparison I also look at close substitutes such as charge cards, credit cards, prepaid cards, and even cash, checks, and barter. I examine the bundle of transactional services payment cards provide. My goal, in large part, is to assess the competitive effects[1] of the debit card interchange fee cap the U.S. Federal Reserve imposed under the Durbin Amendment to the Dodd-Frank Act (2011).[2] In addition to a binding fee cap on banks with net assets in excess of ten billion dollars, it mandated a change in the way the fee is metered.[3] A maximum per transaction fee of 20 cents, binding for most transactions, replaced a typical two-percent negotiated fee. Both of these changes can be expected to have threatened system participants with rent dissipation. I identify the pre- and post-cap equilibrium bundles of transactional services and hypothesize that their differences reflect the parties' cooperative attempts to reduce rent dissipation resulting from the mandatory cap as well as the mandatory change in metering.

One implication I test is that the cap caused or contributed to a decline in the willingness of payment card intermediaries to invest in security, possibly increasing the system's vulnerabilities to the kind of data breaches that have become ever more commonplace. My basic intuition is simple: if consumers of a good—in this case a negotiated bundle of transactional services—cannot assess quality at the point of sale, a regulated price ceiling will likely lead to a reduction in the quality of transactional services.[4] Card security is surely a difficult-to-observe component of the bundle and is therefore a viable candidate for underinvestment.[5]

---

[1] By "competitive" I mean that the parties involved have adjusted the terms of their transactions to minimize deadweight losses in the form of welfare triangles to the extent allowed by the costs of transacting.

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

[3] Note below that the change would be revenue neutral for a ten dollar transaction. Use this as the baseline for explaining the pure metering effect.

[4] Components of quality include but are not limited to point-of-sale services and amenities, warrantees, return policies merchants provide their customers, lines of credit acquiring banks provide their merchants, bundled-in account services issuing banks provide their cardholders, and the range of services card associations such as Visa and American Express provide to issuing banks and to cardholders.

[5] To the extent transactional intermediaries reduce their provision of, say, card security, cardholders and merchants will feel compelled to do so to some extent. Underinvestment ultimately unravels to cardholders

The economics literature on payment card systems has blossomed in the past decade, but it began decades before with William Baxter's seminal work *Bank Interchange of Transactional Paper: Legal and Economic Perspectives* (1983). His paper was the first to propose that payment systems constitute "two-sided markets"[6] because they include transactional services, a conceptually distinct good, tied to the exchange of real goods. The essential attribute of two-sided markets is that they coordinate buyers and sellers of real goods as demanders also of transactional services.[7] One-sided markets are said to coordinate buyers and sellers of real goods only, most fundamentally in a barter system. But, in any system more complex than barter, both buyer and seller stand, in addition, as consumers of the transaction itself, normally facilitated by intermediaries. The transaction may be the mere pairing of buyers and sellers, as in auction and spot markets, or the processing and underwriting of payments, as in the payment card example.

Bank checking is a simple setting in which to illustrate the economics of two-sided markets. A consumer hands a check to a merchant, in most cases tendering it as consideration for immediate possession of the good. The merchant then presents the check to its bank for processing. The merchant's bank finally presents it to the consumers' bank for payment and credits the merchant's account at a point in time depending on the presence and terms of a line of credit secured by the merchant's accounts receivable. The consumer's bank, having reliable information about its own customers, as well as many merchants and their banks, normally honors the draft and later debits the consumer's account accordingly. Early on, banks imposed a fee, or discount, on the funds it remitted to the payee to cover the costs of processing, with the discount tending to diminish over time. In the U.S., for example, banks now honor one another's customers' checks at par.

In Baxter's view, bank checking is a two-sided market because the buyer and the merchant are joint demanders of transactional services, including check processing,

---

and merchants relative inefficiency at providing, say, card security compared to transactional intermediaries.

[6] Others, including Evans and Schmalensee (1995) and Rochet and Tirole (2002), have since formalized the analytics behind two-sided markets.

[7] The Uber app is a salient example.

security, and payment underwriting.[8] One of his main points is that the transaction is nonrivalrous as between the buyer and seller and is therefore a joint product in the sense that it takes both parties to "cause" the transaction to occur. As with all nonrivalrous goods, the merchant's and consumer's demands for the transaction must be summed vertically rather than horizontally to identify the combined value of transacting.[9]

In the context of card payment systems, the buyer/cardholder and the seller/merchant each consume or receive the benefits of the transaction,[10] including payment processing, security, and underwriting. The merchant is said to "pay" the transaction costs in the form of what is called the merchant discount, which in a competitive market should equal the price of the good minus the cost of transactional services. If the cardholder pays $100, normally the merchant's account is credited with about $98, with much of the difference going to the bank that issues the card to the holder, monitors security, and underwrites the payment process by guaranteeing the cardholder's creditworthiness. It appears cardholders pay very little of the two percent transaction cost. The issuing bank may actually subsidize their card usage with various account amenities such as cash or in-kind rewards, unpriced float, free checking, etc., and cover its costs with the merchant discount.[11] Under competition, these adjustments are necessary to equalize the elasticities of demand for transacting between cardholders and merchants.

Whether it is inevitably true or not, many two-sided markets appear to exhibit network effects (Klein, Lerner, Murphy, and Plache, 2014). The more merchants that accept the card its holder carries, the better off the cardholder is. The more cardholders carrying a card the merchant accepts the better off the merchant is. The optimal solution appears to be one in which the card association coordinates the network to clear the market, sets the terms of trade with the merchant and its bank, and allocates the merchant

[8] Baxter seems to recognize that "transaction services" is both a loosely defined category and a multi-dimension one as well.
[9] See Samuelson's formalization. Who does Baxter cite?
[10] What if the "buyer" is an intermediary? Many retailers use credit cards to fund their purchase of inventory, which they then sell to final consumers compensated by card payment.
[11] Some scholars have suggested that the marginal cost of transactional services may be less than two percent, meaning that the merchant ends up subsidizing the cardholders' receipt of rewards and other amenities from their issuing banks (Zywicki, ?). It is plausible that the merchant benefits from these amenities. Airlines are merchants, taking payment from travelers by credit cards. But airlines are also the beneficiaries of issuing banks' travel rewards programs.

discount between various intermediaries involved in processing, security, and underwriting. Payment processing involves any number of steps in a specialized network supported by contractual and customary understandings about what share or strip of the total interchange fee each intermediary will receive, among other things. The card association, whether Visa, Mastercard, American Express, or another has central authority to engage in Coasean bargaining, enforcing and varying the rules to provide system participants with optimal incentives and to earn the resulting residual (Alchian & Demsetz, 1972; Klein, Lerner, Murphy, and Plache, 2014). Since this and other functions along the transactional pathway are noncontractible, the parties rely on reputation and repeat dealing to establish the necessary trust. The brandname reputation of the card association is a substantial part of its capital stock, in this case consisting largely of the residuals it stands to earn in perpetuity from efficient oversight. The reverse is true as well. Inefficient oversight will reduce the wealth of the card association, perhaps to zero. Over time, the market no doubt selects in favor of card associations that efficiently steward their reputations.

The accepted analytics of two-sided markets map nicely with the analytics of any one-sided market in which transacting is recognized to be costly. This means whenever the researcher has a basis for making predictions about the effect of transaction costs on price, quantity, and other terms of trade (including the various attributes of the good being transacted, whether the underlying good or the transactional services necessary to support its exchange). In a so-called frictionless barter system transaction costs are assumed away and the market is, by definition, one sided. In a barter system with frictions, the parties vertically integrate the associated transactional services by performing them personally (see, e.g., Alchian, 1977). This market is two sided but its two-sided quality is masked by vertical integration of transactional services into the exchange of the underlying good. The two-sidedness of the market becomes evident only when specialized intermediaries emerge to perform transactional services. In both settings, the researcher is essentially assessing the influence of transaction costs on the attributes of the exchange.[12]

[12] It seems clear that, except in the frictionless model, transactional services can never be fully vertically disintegrated from the underlying good, which, among other things, should cause puzzlement over exactly

What Baxter coined as two-sided markets more generally reflect the evolution of specialized intermediaries who "own" the potential variability—the residual—in transaction values at various links in the transactional services pathway and across multiple states of the world. Whether the distinction between one- and two-sided markets is a difference in degree or kind, the model Baxter lays out provides important insights into the cost of transacting in any market, from barter to electronic impulses.

As foundation, Part II reviews some of the early and subsequent transaction cost economics literature. Coase's (1937, 1960) seminal works on the nature of the firm and on social costs provide the foundation, but useful insights into the payment card system also come from Demsetz (1968) on the demand for immediacy in transacting, Hirshleifer (1971) on excess search in transacting, Alchian and Demsetz (1972) on the entrepreneurial function, Cheung (1974) on price controls, Barzel (1976) on transaction taxes, Alchian (1977) on money, Klein and Leffler (1981) on quality assurance, and Klein, Crawford, and Alchian (1978) on appropriable rents. Section A of Part III describes the payment card system in greater detail and shows how it fits with the economics of two-sided markets. In Section B, I argue that the parties' demands for transacting are simply derived demands reflecting their respective consumer and producer surpluses from trade in the underlying good. This is because they can be presumed willing to incur transaction costs to facilitate a trade up to the difference between the value they receive from the trade and the value they must forgo. With this foundation, I explain the likely first- and second-order effects of the fee cap and other restrictions the Durbin Amendment and Federal Reserve enabling rules imposed on debit card transactions. Part IV provides a brief summary and concluding remarks.

II. The Transaction Economics Cost Literature

what can be meant by the "underlying good." Even if apples were to float freely in the ether, it would take an individual's time and effort to choose between them. It may be impractical to call the cost of choosing in an individual setting a "transaction cost. With trade, transacting occurs and transaction costs limit exchange compared to a frictionless world, but trade will occur only to the extent that it reduces the individual's cost of choosing below what he faces in the no-trade world. In this sense, the definition that transactions costs are all those costs that do not exist in a Robinson-Crusoe economy becomes a little hazy in a global equilibrium sense. It can be salvaged by restricting analysis to the comparative statics of organizational choice in local equilibrium.

The total cost of payment card exchange is a transaction cost, and one currently subject to considerable specialization. My review of the economics literature on transaction costs starts at the beginning but then proceeds in what I consider logical rather than chronological order of publication date. The literature began with Coase's early (1937) work on the theory of the firm. He posits that using prices to meter market exchange is costly. The firm arises to economize on the costs of using prices. We can think of these costs as transaction costs. 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 are costly.[13] By organizing activity within the firm rather than in the market some transaction costs can be avoided. Exchange within the firm under the direction of the entrepreneur is costly, however, even though it avoids some amount of metering. These costs can also be considered transaction costs. The extent of the firm is determined by balancing the transaction costs of internal, un-priced, 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. Within a single firm, American Express integrates system coordination, payment processing, card issuance, card security, underwriting, and merchant acquisition. The only revealed price is the merchant discount. For my purposes, 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. Roughly speaking, the merchant pays a discount for exchange processing, and this discount is shared between various separate firms in the network chain. For example, Visa acts as the central contracting agent (Alchian and Demetz, 1972), branding cards, coordinating and metering fee allocations, penalizing participants, paying incentives, and establishing security standards. The acquiring banks aggregate and underwrite transactions and provide conditional security, and the issuing banks underwrite cards and also provide conditional security. More specifically, other parties participate in the system as front-end and back-end network aggregators and processers and receive a

[13] What is more, by metering exchange through prices one or both of the parties might reveal entrepreneurial secrets that reduce their ability to capture returns on invested capital (Johnsen, 2001).

compensating portion of the merchant discount for their troubles. Perhaps what is "open" about the open-loop system is that the participating firms have no exclusive role, but instead face competition from rival firms.

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. With no frictions in the market, economic efficiency rather than initial ownership positions determine resource allocation.

Coase never used or promoted the term Coase Theorem, nor did he consider a world of zero transaction costs a coherent benchmark for economic analysis.[14] 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 by way of Coasean bargaining. Markets and other economic institutions exist to capture these opportunities, but the cost of transacting gets in the way. If a blackboard 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 under which they trade allows the uninformed party to trust the informed party. The uninformed party can economize on the cost of becoming informed while expanding his trade with the informed party. Joint surplus rises because the seller's brandname, a nonrivalrous good, saves consumers from having to inefficiently duplicate information. To observe a persistent information asymmetry is to observe a successful market at work.

[14] Coase (2012), at 174. The term is attributed to George Stigler.

Alchian's exploration, *Why Money?* (1978), demonstrates this point and provides important insights into payment systems, one example of which is money itself. The article shows that, unlike traditional wisdom, money is neither necessary nor sufficient to ensure the "double coincidence of wants,"[15] but only to ensure the "double coincidence of information." He imagines 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. Absent money, novices must incur transaction costs to assess the quality of any good they might barter with another novice, which substantially limits joint surplus. A novice can barter with an expert in the expert's good at lower transaction costs under the assumption that the expert can easily assess the value of the good and that he can be trusted to convey this value accurately to buyers and sellers. Experts are assumed to face virtually nil costs of trading with one another in their own goods because both can credibly assure the other of the quality of his goods. Information asymmetry is good.

Now assume novices and experts can evaluate cash at relatively low cost, although experts retain a slight advantage. If the diamond novice wants wheat his most obvious choice is to trade directly with a wheat expert. Since the wheat expert must value the diamonds, for which he has no special expertise, much of the value of the transaction would be dissipated. The diamond novice could sell to the cash expert, then take his cash and buy wheat. This would solve the double coincidence of wants, but it requires the cash expert to incur the cost of assessing the quality of the diamond novice's diamonds. The diamond novice bears this cost in the form of a discount on his diamonds.[16] He must then take the cash and buy wheat from the wheat expert, incurring additional, though modest, transaction costs. As Alchian explains it, "going from a diamond novice through a *C* novice—or even a *C* expert—rather than through a diamond expert first won't help. Some buyer of the novice's diamonds still has to value them.

[15] The double coincidence of wants describes the coordination problem that is thought to arise in the absence of money. Suppose a diamond seller wants to use his diamonds to buy, say, wheat. It may be that the wheat seller does not want diamonds, but instead wants oil. The diamond seller therefore has to find an oil seller who wants diamonds, accomplish that trade, and then go to the wheat seller to trade shoes for wheat. At the very least this requires two transactions rather than one, but it may require more than two. The diamond seller might have to trade diamonds for oil, oil for shoes, and finally shoes for wheat.

[16] Alchian assumes transaction costs also would be higher than if the diamond novice went straight to the wheat expert to trade because it requires an additional transaction.

Evaluation by anyone other than a diamond expert (who becomes a specialist middleman) won't reduce costs."[17]

Because the diamond expert can assess cash at low cost and can be trusted to convey this information accurately, the diamond novice is better off going directly to the diamond expert and selling his diamonds for cash (which the diamond expert gets at low cost from the cash expert), then taking 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."[18] Being a universally low-assessment-cost good, money ensures the double coincidence of information. Access to money, together with the ability to use it to trade real goods with multiple specialists at low cost, increases social surplus and thereby 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 Alchian's simplified economy. When novices go to barter real goods, they are also mutual demanders of "the transaction," just as in two-sided markets, but without intermediaries 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 if his costs of assessing oil are sufficiently low (along with the costs of growing his own wheat), 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 assessment costs he jointly incurs with the contra party. This calculus also applies to the contra party. Barter strikes me as a two-sided market that performs the same balancing act as any two-sided market, with the difference being that there are no transactional specialists in the barter market and no explicit pricing of the transaction for the world to observe. The balancing function is vertically integrated into to what appears to be the simple exchange of real goods. This is not a criticism of the literature on two-sided

[17] Alchian (1978, at 137).
[18] Alchian (1978, at 138).

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 unpacked the multi-dimensionality of trade in real goods. Barzel's (1976) work on taxation provides a compelling example of the second-order effects of taxes on the characteristics of the good. It describes the incentives consumers and producers have 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 tax burden by reducing the number of packs traded. The legal definition of the good being taxed, "packs," does not fully encompass the multi-dimensional nature of value and cost involved in transacting the underlying economic good. Before the tax, consumers and producers had already cooperated to maximize the gains from trade by optimally balancing values and costs on multiple dimensions to arrive at the pre-tax components of the bundle known as the "pack." During the early 1970s, this led to trade in standard packs of 20 cigarettes, each 88 millimeters long, of given quality, with timely delivery to the merchant, in packaging somewhat suitable to preserve freshness, etc., etc., etc.

With imposition of the tax, to reduce the tax burden consumers and producers had an incentive to put more tobacco—or more smoking pleasure—into every pack. Although the new pack increased the per pack marginal value to consumers, marginal cost to producers, and price, 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.[19] The new characteristics of the pack, which were sub-optimal pre-tax, were then conditionally optimal, in essence allowing other attributes of the good besides price and quantity to adjust to establish a new equilibrium. This is exactly the kind of balancing the model of two-sided markets is designed to illustrate.

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, with more

[19] 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.

valuable attributes bundled in, the price of the new packs would be higher than the price of the old packs.[20] With *ad velorem* taxes, this adjustment would increase the total dollar tax per pack, negating the possibility of tax savings. How can the parties adjust to an *ad velorem* tax to reduce the tax burden? 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 previously 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.[21]

In addition to showing how transacting parties can reduce the tax burden by changing the attributes of the goods (real or transactional), Barzel's analysis of taxes further shows 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, or taxes, 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 them they will find opportunities for value capture, which, depending on the setting, may increase or reduce rent dissipation.[22] The tax example provides a simple though compelling demonstration.

Private wealth maximizing parties recognize the ambitions of their counterparts and will seek to pre-empt any opportunities for wealth capture through organizational innovation if doing so generates benefits in excess of transaction costs. Their attempt to

[20] Barzel shows that, owing to the second-order effect of per unit taxes on the characteristics of the good, the tax inclusive price of the good can increase by more than the amount of the tax.
[21] While it is true that airfares may not have fallen in absolute terms, it is sufficient that they increased by less than otherwise.
[22] Passengers might overload the carry-on capacity of the aircraft and carriers might cancel poorly subscribed flights under the guise of mechanical failure.

reduce rent dissipation is exactly what allows economists to explain the form of organization they 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. Part III shows how the Durbin Amendment's regulatory ceiling on interchange fees might have affected payment system security and cardholder welfare, among other things. It also identifies the likely effects of mandating per unit interchange fees where *ad velorem* pricing had apparently been the primary method of metering pre-Durbin. I call these the price control and metering effects, respectively.

One last note on taxes; it is well established in the economics of taxation that who formally pays the tax—whether buyer or seller—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, all else being equal, 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 to what extent, merchants bear the entire burden is an open empirical question. The answer depends in part on the effect on merchants' net sales revenue of accepting a given card. 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 provides all parties with benefits while requiring all parties to bear certain 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 relative to checks, cash, or barter. Having to pay a price, however metered, is better than not getting the good at all. In this sense the tax analysis is inapt for assessing payment card fees.

One of two considerations Alchian assumes away in his treatment of information asymmetries in his simple market 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 dissipate rents 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 with some average quality and prices them accordingly by the pound. 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 overlapping relative valuations for various apple characteristics, the early shopper will buy a sample of better-than-average quality, leaving the remaining apples of lower-than-average-quality. As the day progresses, the grocer must either lower the price to reflect the remaining average quality or face the threat of apple spoilage. Between each price-adjustment, early shoppers gain at the expense of late shoppers. Shoppers are likely to rush to the grocery store. Not only do shoppers incur real costs when they rush, but the grocer's price adjustments incur real costs. Both dissipate rents. It is the equivalent of spending a buck to transfer a buck from one pocket to the other.

Because the parties will adjust downward what they pay the grocer, the costs of sorting feed into reducing the grocer's profit and, in the long run, increase the average price he must charge 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 would be better off. Picking-and-choosing by customers is a cost of transacting to be avoided if possible.

In 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, and as a result there will always be exploitable valuation errors in pricing. 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 from being able to exploit pricing errors.

There are various constraints the parties might adopt to minimize the resulting dissipation. The grocer might, for example, bundle apples together randomly in bags, thereby limiting shoppers' ability to pick and choose and co-opting some measure of dissipating pre-purchase inspection.[23] Sellers might also provide a limited warranty that reduces consumers' incentive to inspect ex ante by promising ex post adjustments in the terms of trade. Finally, sellers might inhibit customers' ability to profit from picking-and-choosing by limiting the quantity of the good they can buy at any one time. In the absence of such a constraint, the first buyer of apples in the grocery store each morning could engage in arbitrage by buying the best apples and then selling them outside the grocery store at a premium price. The arbitrage profits are likely to be small with quantity limitations. Lest this example seem fanciful, consider the enmity concert promoters have for ticket scalpers and the limit they routinely place on the number of tickets available to any given buyer.

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 issued by different originating banks. 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 to some extent it should be in participants' joint interest to avoid this dissipating activity. It is in part 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 magically 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. In the economics literature, such goods are known as experience goods. The problem with experience

[23] During a recent trip to the grocery store, I found mesh bags filled with a variety of apples as well as oranges. It was clear the grocer had done the bagging.

goods is that the seller might promise to provide high-quality units of the good at a price sufficient to cover the necessary costs, while secretly cutting quality to reduce those 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 solution is illustrated in Figure 1. For simplicity, assume unit sales revenue is strictly a function of time, shown on the horizontal axis. Value, cost, and price per unit of the good 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 will refuse from the start to pay any price higher than P_L. Trade in the high-quality good would never occur, even though trade could hypothetically increase the social surplus. A so-called "lemons" market would prevail (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 1. 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 plus the double 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 ∞, reflected in the shaded area. This flow is a perpetuity whose present value at Time 1 (assuming Time 2 is the moment the producer would be caught cheating) is $[P^* - MC_H]/r$, where r is the appropriate discount rate.[24] 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 a high-quality

[24] This perpetuity includes the double cross-hatched box, which the seller will capture regardless of whether or not he cheats.

good. As in any consumer advocacy setting, mandating lower interchange fees on debit transactions is unlikely to help consumers if the good in question is an experience good.[25]

This does not quite end the story because it may leave the producer with a surplus, which cannot persist in a 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 specialized 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 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.

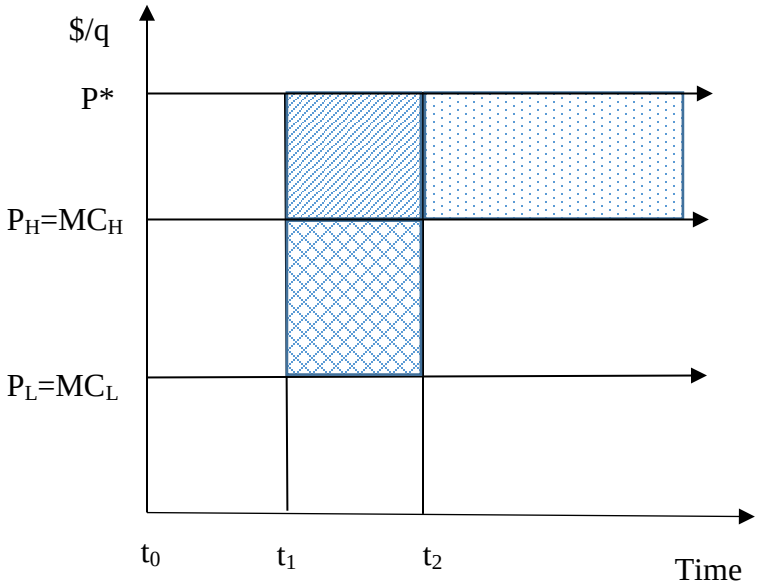

Figure 1: *Quality Assurance*

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

The card association's brandname is a specific asset that bonds the quality of payment processing, including card security. Because of this, however, it is subject to opportunistic appropriation by others in the system. Klein, Crawford, and Alchian (1978) address the problem of appropriable quasi-rent. A quasi-rent is the promised payment to a specific asset necessary to cover the opportunity costs of putting it in place, but once in place its opportunity cost declines dramatically and the promisor can attempt to appropriate the quasi-rents by opportunistically reducing the offer price. If contract enforcement is costly, the promisor may succeed in his demands. The authors show that vertical integration (among other organizational forms) can be used to solve this problem.

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 their trades, 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 given moment that there will be a contra party in the market willing to take the other side of the trade at a competitive price and in the same volume?

NYSE specialists perform this function. They receive 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 their knowledge of aggregated limit orders in the stock or stocks they cover, they post the price at which they are willing to buy—the bid price—and the price at which they are willing to sell—the ask price. The bid-ask spread at any moment is the difference between the ask price and the bid price. As the day progresses specialists adjust their bid-ask spread according to the arriving limit orders and either facilitate the trades of floor brokers or trade for their own account. Whenever a specialist buys and sells 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, he typically makes money because of his superior private information. In doing so, 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 policies or existing laws (e.g., lemon laws). 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 revenues find their 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. The customer gets his goods immediately and the merchant will have his account credited by the end of the day without having to bear the costs of float or the cardholder'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. It 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.

III. The Payment Card System as a Two-Sided Market

A. Payment Card System Overview

The textbook 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, checks, or cards, must incur lower transaction costs over some range than barter or the parties would decline to use them. Conditional on the availability of more efficient arrangements, potential gains from trade would be left on the table with barter.

Baxter (1983) traces the evolution of transactional paper in the U.S. from the time of the early national banking system through the development of credit cards. Legal tender at the time of the new republic was scarce. Early forms of transactional paper consisted primarily of bank notes and bank drafts (personal checks). For the buyer of a good or service from a given locality to arrange payment to a distant seller in anything but cash—which had obvious problems of its own—required a byzantine series of transactions involving the buyer's and seller's banks and often one or two corresponding banks. Each bank charged a fee for exchange—the "discount"—along the way that was not necessarily disclosed in advance. 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 at par rather than at a discount. Apparently whatever risks of nonpayment and other adverse events they bore largely balanced out across banks. Moral hazard must have been sufficiently low that the banks could diversify any unsystematic risk.

The use of payment cards 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 merchants such as prominent hotel and restaurant chains to honor the cards,[26] with a guarantee that the bank as underwriter would pay the cardholder's debt. These banks have come to be

[26] Or perhaps the acquiring bank gets its moniker from the fact that it acquires revenue from issuing banks on behalf of the merchant.

characterized as "acquiring" banks. 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 together with modestly delayed payment in exchange for the additional sales to cardholders.

The predecessors of the MasterCard and Visa credit card associations appeared in the 1960s. These associations began as mutuals owned by various regional member banks and were designed 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 2 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 his bank in cooperation with the branded association. The system performs two basic back-office 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.

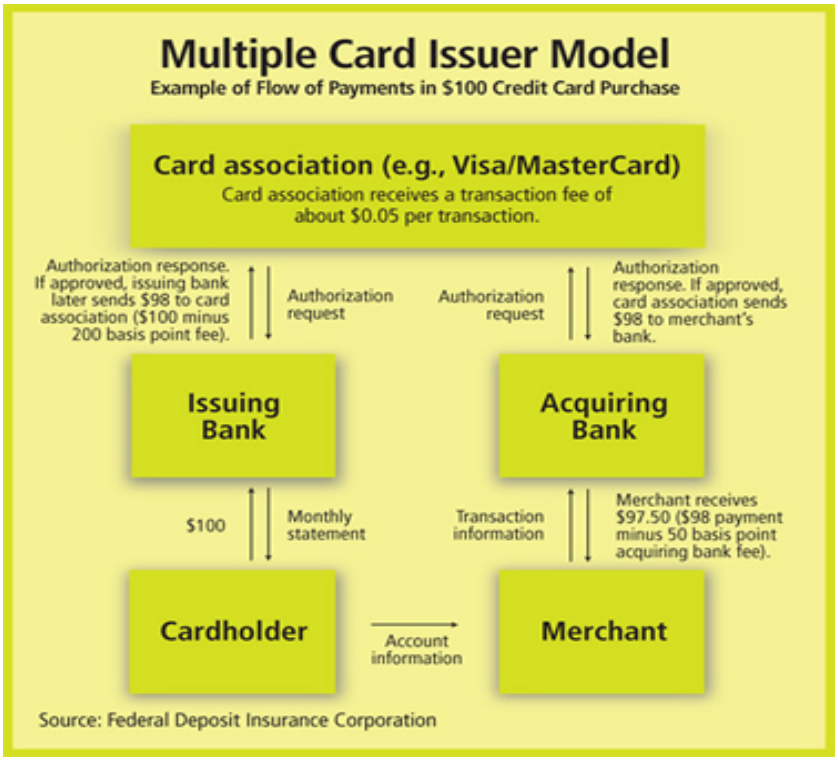

Figure 2: *The Payment Card System*

The open-loop system is considerably more complex than what Figure 2 suggests. Any number of specialized intermediaries, most importantly electronic aggregators and technical networks, 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 timely authorization, aggregation, and clearing and settling functions as described below.[27] What is more, the customer and the merchant get the benefits of immediacy. The customer gets his goods the moment his card is authorized and the merchant gets the funds, often by the end of the business day. The issuing bank makes immediacy possible by the underwriting—or bearing the financial risk of—the transaction.

Authorization: The cardholder presents the association-branded card containing his account information to the merchant at the point of sale. Either the merchant or the cardholder swipes the card into a point-of-sale electronic terminal or a payment gateway

[27] See, e.g., https://www.ippay.com/index.php?q=merchant_processing_overview.

through a secure connection from a website, retail location, or a wireless device.[28] The payment gateway may be an independent firm as well as part of an electronic network. It 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 issuer posts the transactions to the cardholder accounts and sends cardholders a monthly statement.

B. The Economics of Two-sided Markets and the Durbin Fee Cap

Baxter models the demand for and supply of transactional paper as a two-sided market. Transactional paper is written evidence of the amount of the underlying good being transacted, the timing of delivery and payments, the source of certification, warranties and exclusions, provisions for the negotiability of the paper, and other terms. Consumers and merchants jointly demand this amorphous bundle of transactional services, with one unit of the good—the transaction—tied by definition to the purchase and sale of the underlying good or goods. In Figure 3, consumers and merchants each

[28] It appears that brick-and-mortar merchants use both point-of-sale terminals and payment gateways, while online merchants rely exclusively on payment gateways.

have separate demands for transactional services (D_C and D_M), but, as with nonrivalrous goods generally, these demands must be summed vertically rather than horizontally to arrive at total market demand, D_T. Similarly, the total market supply of transactional services, S_T, is roughly the vertical summation of the participating intermediaries' marginal costs because it takes their services jointly to complete a single transaction.[29]

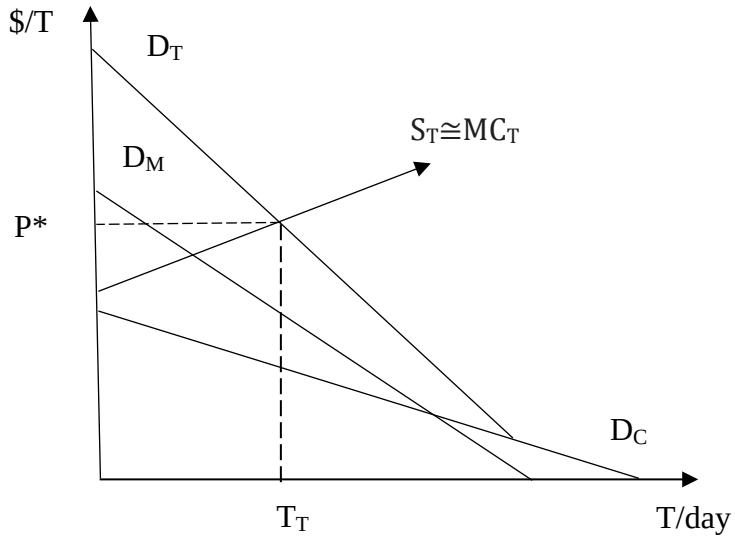

Figure 3: *Baxter's Model of Two-Sided Markets*

Equilibrium establishes the total interchange fee, or merchant discount, which is equal to the aggregate marginal costs incurred by the participating intermediaries. T_T on the horizontal axis shows the total number of transactions, say, per day, processed in the transactional paper market. At T_T, Figure 3 shows that merchants' marginal willingness to pay for transactions, MV_M, exceeds what consumers are willing to pay, MV_C, for all but a very high rate of transactions. Therein lies one of Baxter's main points; there is no reason these amounts should be equal. Instead, the market equalizes the elasticities of demand between consumers and merchants. For given transaction costs, the parties expand trade until the marginal gain in consumer surplus is equal to the marginal gain in producer surplus, namely zero.

[29] For simplicity I depict S_T as a straight line rather that a more realistic marginal cost curve whose slope increases as the rate of transactions increases. At this point the simplification is immaterial.

Figure 4 takes a closer look at Baxter's basic model in the context of payment cards. Panel A starts with the demand for and supply of a standardized real good, D and S. In a no-friction model the equilibrium price of the good is P* and the equilibrium quantity traded per day is Q_{NF}. The sum of consumer and producer surpluses reflect the gains from trade. If transacting is costly, consumers and merchants must either incur their own transaction costs (which cuts into their respective surpluses) to conclude the trade, or they must rely on specialized intermediaries to assist them. To the extent intermediaries can perform transactional services at lower cost, all things considered, the parties delegate that function to them. For each unit of the underlying good consumers and merchants trade they will, by construction, also jointly demand one unit of transactional processing. As such, their joint demand for transactions is derived from the demand for and supply of the underlying good.

These derived demands for transacting are their consumer and producer surpluses for the underlying good. In Panel A, if consumers value the first unit of the underlying good at MV_1 but must pay only P* to buy it, by construction $MV_1 - P^*$ reflects their consumer surplus as well as the additional amount they would be willing to pay for transactional services to trade the first unit. At Q_T consumers value the first unit of the underlying good at exactly P*, and they would be willing to pay nothing for further transactional services to conclude a trade. Panel B shows their derived demand for transactional services as D_C. Looking back to Panel A, merchants are willing to provide the first unit of the good for as little as MC_1 but stand to receive P*. The difference $P^* - MC_1$, is their producer surplus as well as the additional amount they would pay to conclude the first trade. At Q_T they are willing to pay nothing to conclude the marginal trade. Merchants' derived demand for transactional services is shown as D_M in Panel B.

Unlike Baxter's analysis, mine shows that consumers' and merchants' derived demands for transacting must intersect on the horizontal axis, along with the total demand for transacting.[30] In Panel B, the equilibrium number of daily transaction under a negotiated interchange fee is T_{IF^*}, far fewer than the number of daily transactions in the

[30] This assumes that the cost of transactional services is the only friction in the system being examined. Transaction costs the parties bear themselves and taxes would constitute other costs of trading whose account might cause their demands to diverge at the horizontal axis.

no-friction model, T_{NF}. IF* is the equilibrium interchange fee per transaction, while MV_{C*} and MV_{M*} reflect consumers' and merchants' respective marginal valuations at T_{IF*} as well as the portion of IF* they bear.

Returning to Panel A, with Q_F units (corresponding to T_{IF*} transactions) being transacted P_B is the price consumers pay inclusive of transaction costs, and P_A is the price merchants receive net of transaction costs. The difference, $P_B - P_A$, is the interchange fee or merchant discount and is analogous to the standard bid-ask spread in financial markets (see Demsetz, 1968). In practice, this fee would be roughly two percent of P_B absent the Durbin fee cap. Consumers' and merchant's respective surpluses are obviously much lower than in the frictionless model, as is the quantity of the good traded. The shaded triangle reflects gains from trade the parties capture in the frictionless textbook model that are not worth capturing when transacting is costly.

This analysis exactly parallels the standard economic analysis of the first-order effects of taxation, but rather than paying a tax to the government determined by fiat the parties pay a negotiated interchange fee to participating intermediaries. In Panel A of Figure 4, S' shows the total supply of the underlying good buyers perceive including one unit of (ill defined) transactional services, as the vertical sum of S and S_T. Recall that Barzel (1976) distinguishes between two forms of taxation to assess second-order effects: 1) a fixed dollar tax on each unit of the good traded, called a per unit tax and 2) a percentage tax whose dollar amount increases with the dollar value of the transaction, called an *ad valorem* tax. As far as first-order effects are concerned the analysis of per unit and *ad valorem* taxes is virtually identical. With either tax the amount of the good traded declines, the price to the buyer rises, the price to the seller net of the tax falls, the tax burden comes out of the parties' respective surpluses depending on the relative elasticities of demand and supply, and in addition the parties suffer the standard deadweight loss equal to the shaded triangle. The only difference is that with a tax imposed, say, on the merchant, the supply curve the consumer perceives inclusive of a per unit tax shifts parallel to S but the supply curve the consumer perceives inclusive of an *ad velorem* tax rotates around the (here shifted) origin of S, as in Panel A. There is invariably some per unit tax whose first order effects will be identical to a given *ad velorem* tax.

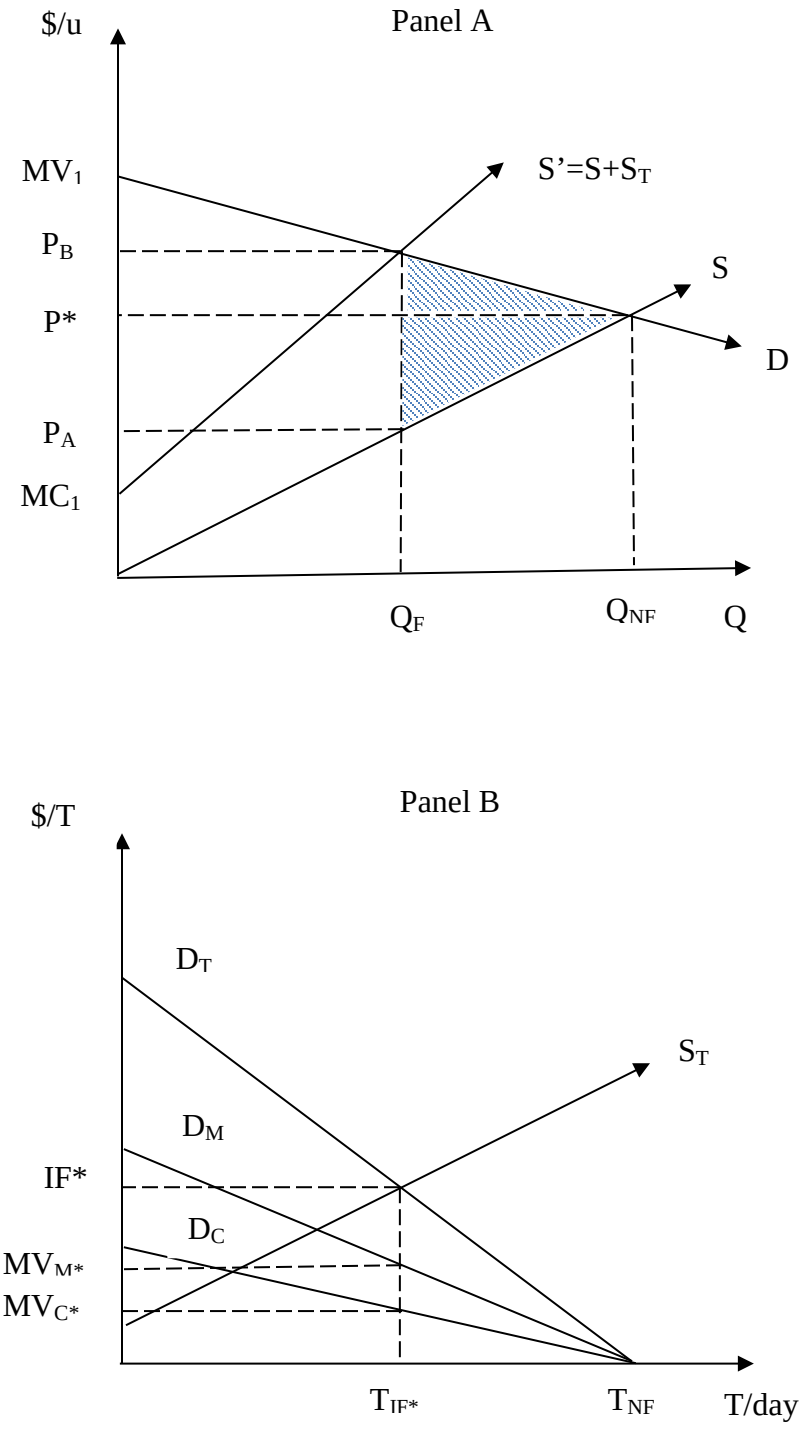

Figure 4: *A Closer Look at Two-Sided Markets*

All but the most rudimentary economic goods generate value on multiple dimensions that the parties cannot costlessly identify—especially the tax collector—at the point of sale. This is one of Baxter's basic points. Unless the tax collector carefully meters and controls these dimensions of value the parties can manipulate them to reduce their joint tax burden. Recall the example from Part II describing an unfettered market for filtered cigarettes sold by the pack. Suppose, initially, the price, P*, is five dollars per pack. Suppose the tax collector then imposes a tax of one dollar per pack on all trade in cigarettes and effectively enforces it. According to a first-order analysis the per unit tax will increase the tax-inclusive price to consumers and reduce the net-of-tax price sellers receive, total trade in cigarettes, and gains from trade reflected by the standard welfare triangle. How much the tax-inclusive price to consumers rises and how much the net-of-tax price sellers receive falls depends on the relative demand and supply elasticities for packs of cigarettes. Under plausible elasticity assumptions, the price of the good will rise by one dollar at most but more likely by less because of the parties' economizing adjustments to the tax.

What about the second-order effects of a per unit tax on cigarettes? With the first-order tax analysis the unit of the good is assumed fixed in all its dimensions; a pack is a pack is a pack. But in the real world the parties can vary the valuable dimensions of the pack in any number of ways that minimize tax payments. From their standpoint the tax is a rent subject to (re-)capture. They will find it in their mutual interest to add in attributes of the good whose value exceeds the cost of bundling but reduces the number of packs traded as well as the rents paid to the taxman. Pre-tax, the bundle was optimal in this regard, but post tax it is not. Higher quality tobacco can be used in manufacturing, more tobacco can be put into the cigarettes (100mm rather than 88mm cigarettes), packaging can be improved to better ensure freshness, cigarettes' can be rushed to the retail counter with less delay, and advertising can be increased or adjusted to provide greater value to smokers via brand identification. These are just a few of the many ways buyers and sellers can cooperate to reduce the burden of the per unit tax. In essence, the new pack will contain more smoking pleasure than the old pack, its price inclusive of the tax will undoubtedly rise, fewer packs will be traded, and the total tax burden per unit of smoking pleasure will fall as will the forgone gains from trade. The parties economizing

behavior allows them to avoid transferring rents to the government in the form of tax receipts. Barzel shows that these second order effects can be sufficiently large that the price of the good increases by more than the tax, an impossible outcome under a first-order analysis.

Barzel's analysis can be applied to payment cards to identify the competitive effects of the Durbin Amendment's fee cap on debit card transactions. Conceptually, the fee cap can be decomposed into two binding constraints on consumers and merchants. First, at least for transactions larger than ten dollars (two percent of which equals Durbin's 20 cent per transaction fee cap), it imposed a price ceiling on the interchange fee. Second, it changed the method of metering the fee from a percentage basis (two percent) to a per transaction basis (20 cents). I focus initially on the metering constraint by assuming all transactions, pre-cap, happened to be for ten dollars. This allows me to isolate the effect of moving from a percentage fee to a fixed dollar fee per transaction.

With imposition of the per transaction fee cap consumers and merchants would have found it in their mutual interest, at the margin, to add more valuable attributes into the transaction to the extent doing so would allow them to reduce total interchange fees by more than the added cost of bundling. This could include changes in the underlying goods over time and/or changes in the associated transactional services. At the margin, for example, merchants and consumers would now have a shared interest in increasing the durability of capital goods. By way of example, longer lasting razors would cost more but would need to be replaced less often, thereby avoiding future interchange fees. Merchants might also find it worthwhile to provide more information about their goods to consumers at the point of sale through better-educated salespeople, with the price of the goods increasing to cover the added cost. To the extent information about how to properly use the good is a substitute for the good itself, interchange fees would be avoided because it would lead to fewer transactions.[31]

[31] The buyer of a simple flat metal file who does not know that using it in both directions (back-and-forth) will ruin it will end up having to replace it more often than the buyer who is instructed to use it properly, in only the forward direction. No doubt other examples likely adjustments be identified. Note, I am making no statement about what people actually cognize, only about what increases their chances of survival in a competitive environment (Alchian, 1950).

The most obvious adjustment consumers and merchants might make would be to aggregate what would otherwise be multiple transactions into a single transaction. Consumers will prefer razors in packs of 10 over packs of five. Rather than buying different types of goods from various specialty merchants in multiple transactions, consumers will tend to use their debit cards to buy from big-box variety stores, so-called "one-stop shopping." The number of transactions subject to the tax would then decline. Big box variety stores would gain market share compared to their mom-and-pop rivals. This effect would likely undercut to some extent the durability and informational effects discussed above.

The parties might find ways to impose greater responsibility for security and other functions on transactional intermediaries. To reduce rent dissipation under the negotiated two percent fee, merchants and consumers would have found it mutually agreeable to undertake certain security precautions (and other functions) themselves, that is, to vertically integrate to some extent rather than being charged a higher percentage interchange fee to cover security precautions various intermediaries might otherwise provide. Under a per transaction fee they would likely attempt to shift some of these functions to transactional intermediaries,[32] that is, to bundle the costs of added security precautions into the merchant discount. Unlike in the tax context, such shifting may be difficult in the interchange fee context because transactional intermediaries know their business and have contrary interests of their own to press unless compensated for bearing added costs. It is worth noting that, in contrast to security issues, the adjustments to durability, point-of-sale information, and transaction aggregation discussed above would seem to be a matter of indifference to transactional intermediaries.[33]

It should be obvious that these adjustments will increase the dollar amount of the average transaction to something in excess of ten dollars. In an initial world of uniform ten dollar transactions, the price ceiling constraint of the fee-cap would begin to bind if it did not do so from the outset. Alternatively,—relaxing the uniform $10 transaction assumption—if some transactions were above and some below $10, it would bind for the

[32] Consumers might begin relying on their issuing bank for minimum balance alerts rather than closely monitoring their accounts. Merchants might . . . (?).
[33] In fact, issuing banks might prefer larger but fewer independent transactions concentrated with relatively few merchants.

larger transactions from the outset. Standard first-order analysis of price ceilings merely says that some observable (legal) price ceiling will increase the quantity demanded and reduce the quantity supplied, leaving excess demand and consumers and merchants willing to pay much more for the traditional bundle of transactional services than before. Second-order analysis of price ceilings suggests that the buyer and seller as cooperating parties will try to reduce rent dissipation by stripping out valuable attributes of the good rather than building more such attributes in, thus muting or possibly overwhelming the effects of the per unit metering constraint.

Figure 5 depicts the first-order effects of a pure interchange fee ceiling in the debit card transactional service market. It shows only the total demand for transacting by consumers and merchants and the total supply, $S_T \cong MC_T$, as marginal cost aggregated across all intermediaries. In this case I assume the fee cap leaves the method of metering the fee unchanged. For example, assume it mandates that the maximum fee is one percent rather than the typical two percent negotiated fee.

As before, the equilibrium negotiated fee is IF*, or two percent of the total transaction price (not shown), and intermediaries perform T_{IF*} transactions per day as determined by the intersection of S_T and D_T. Also intersecting at this point is a short straight line labeled Line 2% that shows how the interchange fee would adjust as the number of transactions varies around the equilibrium. In addition, the straight line labeled Line 1% shows how the interchange fee would adjust as the number of transactions varies under the Durbin fee cap regulations. It is a version of Line 2% rotated down by 50 percent because, by assumption, Durbin cut the percentage fee in half. Intermediaries are now limited to charging a fee that meets this constraint.

The regulation limits the fee they can charge but it has nothing to say about the number of transactions they must perform per day. Rational intermediaries' decision rule is to perform no more than what allows them to cover their marginal cost. This occurs at T_S, where Line 1% intersects MC_T. Under the Durbin regulations, IF_{DR} is the new interchange fee of one percent shown on the vertical axis. At this interchange fee, consumer and merchants jointly would like to do T_D transactions per day. T_D minus T_S illustrates the excess demand, or shortage, typically associated with binding price ceilings.

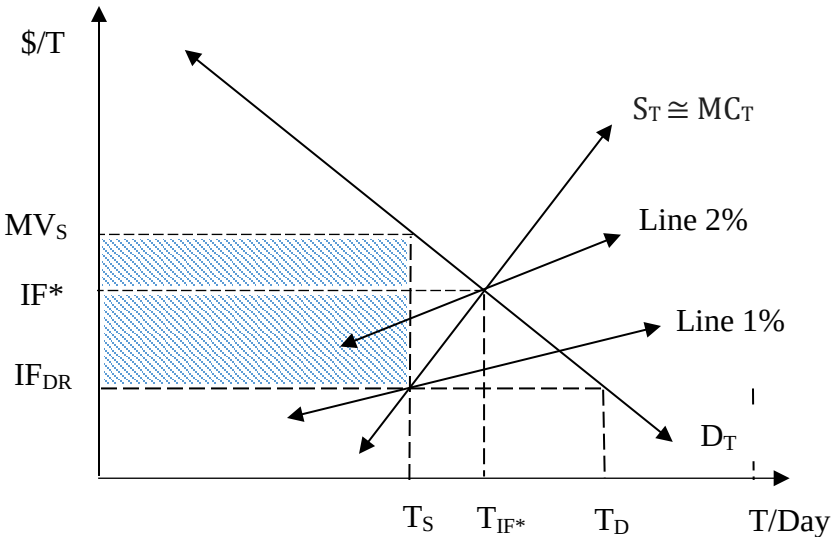

Figure 5: *The Pure Price Control Effect*

As to second-order effects, unless the regulator compels the intermediaries to expand transactional processing beyond T_S some merchants and consumers must go without. Consumers, merchants, and intermediaries will suffer the standard deadweight loss from transactions that could be done at a cost that is less than the value they would generate. What is more, notice that, at T_S, consumers and merchants are jointly willing to pay as much as MV_S for the marginal unit of transactional services, but that by law they are allowed to pay no more than IF_{DR} = 1%. The difference, MV_S minus IF_{DR}, is an economic rent. Summed across all T_S transactions, the total rent is shown by the cross-hatched rectangle. This rent is not exclusively assigned to any specific set of consumers and merchants, who must therefore compete for transactional services in some way other than by offering a higher interchange fee. Some of these rents will undoubtedly be dissipated because open access forces the parties to engage in costly nonprice competition. There are also myriad ways they might limit the rent dissipation but doing so, again, is costly.

Debit card holding consumers might compete by offering to maintain a minimum balance in their debit accounts, by paying a yearly debit card fee, or by accepting reduced spending rewards, that is, by unbundling services. Issuing banks might become disinclined to issue debit cards to consumers with high risk of declined transactions, a history of frequent disputed charges or fraudulent card use, a history of frequent lost cards that must be reissued, etc. They might also underinvest in debit card-specific security (to the extent doing so does not effect credit card security). One way of doing this to raise the number of false negatives in purchase authorization for debit cards, thereby reducing convenience to the consumer and merchant. There is no doubt that one effect of the Durbin regulations has been to leave marginal debit card holders with no card or even no banking services at all (Zywicki, ????). Merchant banks may welcome competitive efforts by merchants as well. They may increase the delay between the time the transactions occurs and the time they credit the merchant's account. Mom-and-pop retailers may be cut from the list of approved vendors. Some of the services formerly provided by intermediaries will likely be vertically integrated by consumers and merchants, and inefficiently so. In essence, the parties will make those competitive adjustments that render T_S transactions for a one-percent fee a new equilibrium. This equilibrium had always been available to the parties, but under competition they considered it suboptimal.

Analysis of the second-order effects of moving from a two percent negotiated fee to a 20 cent per transaction fee cap suggests that the metering effect would lead the parties to bundle more services into the good, including transactional processing, but the price control effect would lead them to unbundle services. It is an empirical question which effect would be expected to dominate. It is possible that the two effects are not always mutually exclusive depending on the attributes of the good and the transactions in question. It may be, for example, that transactional services that are complements to the underlying good will be bundled out, while those that are substitutes will be bundled in, or vice-versa. Empirically testing the predictions of the model is a task that awaits further research.

V. Summary and Concluding Remarks

On its face, the price ceilings 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 card system must be low compared to the alternatives, such as cash or checks.[34] 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.[35] 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 a sound business model. Others have noted the apparent absence of any kind of market failure that might justify regulation (Rochet and Tirole, 2002) or the presence of market power by card associations (Wright & Zywicki?).

Beyond that, the Durbin fee cap applied exclusively to the somewhat transparent open-loop system and not on the relatively opaque closed-loop system. Level heads might wonder why Congress chose to punish transparency. Moreover, placing a price ceiling on a good—an ill-defined bundle of transactional services—whose quality is difficult to assess ex ante is almost sure to lead the parties to seek a new equilibrium bundle whose quality is reduced, quite possibly with the parties that are most efficient at providing security protections, the card associations and issuing banks, shifting some portion of the burden onto consumers and merchants. In this setting, and many others, regulators have failed to keep pace with what is now considered garden variety economic theory. Premium prices for experience goods provide consumers with valuable information about quality that allows them to remain efficiently ignorant of the good's substantive attributes at the point of purchase, in essence allowing them to assess quality after purchase rather than before. As Alchian put it over 35 years ago, "Because most of

[34] Although checks are honored at par within the U.S., this does not mean using them is free of transaction costs. I 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.
[35] Cite Zywicki?

the formal economic models of competition, exchange, and equilibrium have ignored ignorance and lack of costless full and perfect information, many institutions of our economic system, institutions that are productive in creating knowledge more cheaply than otherwise, have been erroneously treated as parasitic appendages."[36]

[36] Alchian (1977), at 140.

LIST OF REFERENCES

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

Armen A. Alchian, *Uncertainty, Evolution, and Economic Theory*, 58 J.P.E., 58 (1950), pp. 211-221.

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

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

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, Michel A. Habib, 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).

Coase, Ronald H., THE FIRM, THE MARKET, AND THE LAW (U. Chicago Press, 1988).

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, Robert G. Crawford, and Armen A. Alchian, *Vertical Integration, Appropriable Rents, and the Competitive Contracting Process*, 21 J. LAW & ECON. (1978), pp. 297-326.

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

Klein, Benjamin, Andres V. Lerner, Kevin M. Murphy and Lacey L. Plache, *Competition in Two-Sided Markets: The Antitrust Economics of Payment Card Interchange Fees*, Antitrust Law Journal, Vol. 73, No. 3 (2006), pp. 571-626

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

Wright, Joshua D., and Todd J. Zywicki, . . . ?

Zywicki, Todd J., *TheEconomics of Payment Card Interchange Fees and the Limits of Regulation* . . . ?