CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

REVISITING WILLIAM BAXTER'S PERSPECTIVES ON BANK INTERCHANGE OF TRANSACTIONAL PAPER

Thomas Brown

O'Melveny & Myers

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

A TRIBUTE, OF SORTS, TO WILLIAM F. BAXTER'S "BANK INTERCHANGE OF TRANSACTIONAL PAPER"

Thomas P. Brown*

ABSTRACT

In 1983, the Assistant Attorney General of the Antitrust Division in the United States Department of Justice, Bill Baxter, did something that would be unfathomable today. He published an academic paper in a scholarly journal that related directly to a piece of antitrust litigation then pending in federal court in which he had served as an expert. The paper did not ignite a storm of controversy. Indeed, outside of the court presiding over the litigation to which Baxter's article related, Baxter's paper attracted little immediate attention. Even twelve years ago, when a group of friends and colleagues gathered to celebrate Baxter's work, this paper took a distant back seat to his tenure at the Department of Justice, his monograph on environmental law, and his one article on choice of law. Today, the paper is recognized as the seminal work on a topic that has attracted considerable attention for the last several years and seems likely to remain on the public agenda in the United States and elsewhere for the indefinite future: interchange.

The consensus on Baxter's paper ends there. There is considerable disagreement about what Baxter's paper actually says. For example, Jean Charles-Rochet & Jean Tirole credit Baxter for observing (1) that the decision to use a payment type requires coordination between the consumer and the merchant, (2) that the merchant and consumer in a four-party payment system may be served by different payment institutions, and (3) that maximization of output frequently requires a transfer from one side of the system to the other.[1] Dennis Carlton extracts a different lesson from Baxter. According to Carlton, Baxter's paper demonstrates that interchange can be used to enable merchants to charge two sets of prices: a higher price for cash customers and a lower price for credit customers.[2]

The various interpretations of "Bank Interchange of Transactional Paper" flow from two omissions in the paper that a contemporary reader will notice—a formal model and discussion of the work of other economists. Baxter's article has none of the former and very little of the later. It precedes by a few years the modeling revolution of Industrial Organization, and like other famous and roughly contemporaneous articles,[3] it makes little effort to explain where it stands in relation to the contributions of other economists. Baxter's article limits its discussion of the work of other economists to two short footnote discussions of an article by Bowen entitled The Interpretation of Voting in the Allocation of Economic Resources and the classic article by Landes and Posner, Market Power in Antitrust Cases.

With some trepidation,[4] this essay attempts to make the going easier. It provides a short map of the paper. It also fills in some of the obvious holes in Baxter's article and flags portions where an unwary reader might get trapped. Baxter's article, like a proverbial Michelin-starred restaurant, is worth the trip. But it is also worth attempting to smooth an otherwise bumpy journey.

* O'Melveny & Myers, U.C. Berkeley Law School. I want to thank Richard Schmalensee and Thomas Hubbard for comments on an earlier draft of this introduction. This paper does not represent the views of O'Melveny & Myers or any of its clients, and the errors and omissions are entirely my own.

I. A TRUNCATED ROAD MAP TO BAXTER'S BANK INTERCHANGE OF TRANSACTIONAL PAPER

Baxter's paper follows a simple outline. It contains three sections labeled as follows: "I. The Theoretical Viewpoint;" "II. The History of Four-Party Transaction Vehicles;" and "III. Conclusion." Like other features of the paper, the apparent simplicity is deceiving. The first and second sections each contain subsections. The first has two—"A. The Demand for Transaction Paper" and "B. The Supply of Transactional Paper." The second has three—"A. The Practice of Paying Checks 'At Par,'" "B. Bank Credit Cards and the Interchange Fee," and "C. Modern Developments." None of the subsections has sub-subsections, though the subsections devoted to "at par" checking and interchange would greatly benefit from them, as they cover quite a bit of ground.

In the interest of brevity, this essay devotes most of its attention to the sections central to Baxter's discussion of interchange—i.e., "I. The Theoretical Viewpoint" and "II. B. Bank Credit Cards and the Interchange Fee." It skips entirely the discussion of "The Practice of Paying Checks 'At Par'" and offers only limited observations about the "Modern Developments."

A) A BRIEF GUIDE TO BAXTER'S "THE THEORETICAL VIEWPOINT"

Baxter's paper does not, at least at the outset, waste any time. After a brief two-paragraph introduction, it jumps into a discussion of four-party payment systems by offering a generic vocabulary to describe those systems. The introduction of this vocabulary plays two important roles for the discussion that follows. First, it literally defines away the obvious differences between checks, credit cards and other forms of non-cash payments that might otherwise complicate the narrative. Second—and this was more important for the case to which this article related than any overt goal of the paper itself—the common vocabulary tends to suggest some degree of interchangeability or substitution among the instruments.

Baxter's vocabulary for four-party payments is quite simple. He posits the following participants:

1) a "merchant (M)" who receives transactional paper in exchange for goods or services;

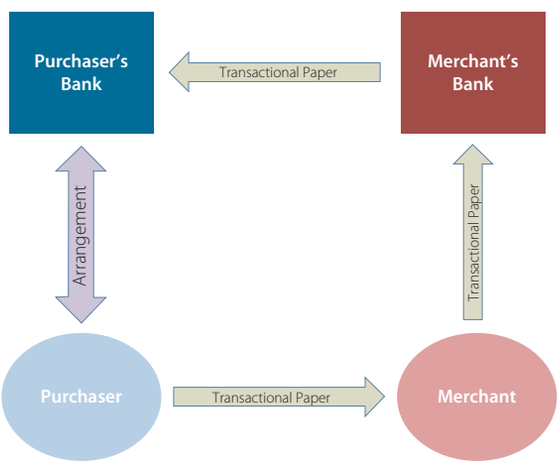
2) a "merchant's bank (M bank)" where M deposits its transactional paper;

3) a "purchaser (P)" who gives M transactional paper in exchange for goods or services; and

4) a "purchaser's bank (P bank)" where P has established "an arrangement that contemplates acceptance of and payment against" the transactional paper presented by P to M.

Baxter's vocabulary, although useful for advancing the points noted above, has one significant drawback. It omits any role for the administrator of the system. In other words, most of the systems that Baxter labels four-party systems are actually five-party systems. This is not as obvious when the system is introduced verbally as Baxter does, but the omission is striking when Baxter's instructions are illustrated.

BAXTER'S FOUR-PARTY TRANSACTION WHITER VISA, MASTERCARD OR THE FED?



Baxter's vocabulary provides the foundation for the paper's first major insight: the selection of a medium of exchange, unlike the decision about whether to purchase a traditional product, is contingent on the choices made by the counter-party to the transaction. Baxter draws the distinction with aid from a pedestrian example. He asserts that when a consumer contemplates purchasing a pair of shoes, the consumer's evaluation of the benefit from those shoes "is usually independent of other consumers' evaluations."

Payments, Baxter claims, are different. In order for a purchaser and merchant to use a particular payment instrument, both have to agree to it:

Rather than considering the demands of P and M as demands for separate products, define one unit of product to consist of the bundle of transactional services that banks must supply to P and M in order to facilitate the execution of one exchange of goods or services between P and M.

Baxter's insight that demand for payments requires coordination among payers and recipients is, as others have observed, profound. But if anything, the paper underplays the significance of the observation by failing to distinguish it from the work of other economists. Long before Baxter wrote his paper, economists had devised tools to model the impact that one person's decision might have on another. Both Alfred Marshall and Arthur Pigou had examined and debated the importance of externalities, and positive as well as negative externalities had appeared in models of everything from pollution to proliferation of intellectual property.[5]

Similarly, the challenge of reaching optimal outcomes through independent action had been a topic of conversation in economic circles at least since John Nash had helped introduce the world to game theory.[6] Even though the works of Marshall, Pigou and Nash do not directly anticipate Baxter's insight, the paper would surely be easier to understand had it taken the time to explain why.

After introducing the vocabulary, Baxter launches into a description of joint demand for transactional services that accompanies Figure 1, a graphical representation of that demand. The graph depicts two crossed demand curves—one for merchants (denoted dM) and one for purchasers (denoted dP)—that are summed "vertically" into an aggregate demand curve denoted d'.

The paper explains that the diagram should be understood to show the relationship between price and quantity for transactions conditioned on P and M coordinating their relative contributions to pay for the jointly consumed service.

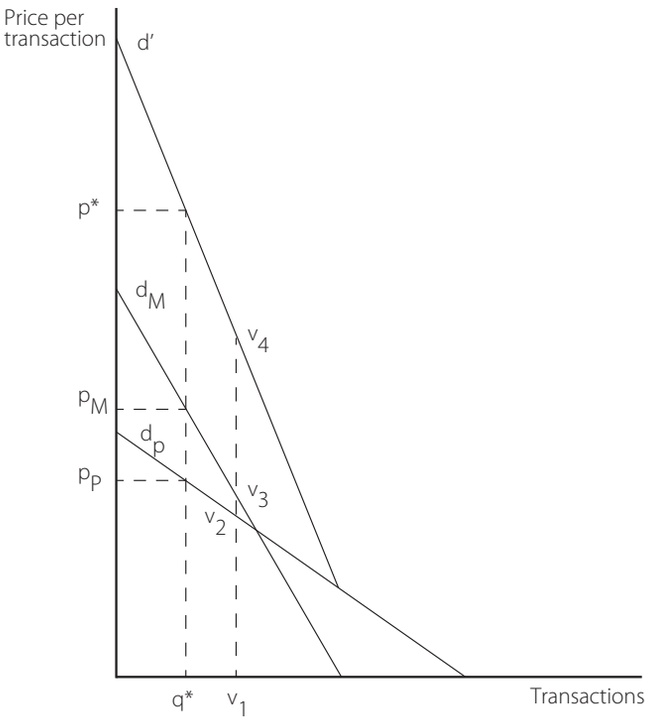


Figure 1

Baxter's paper then takes the discussion to the supply-side. Here, the coordinating parties are P's bank and M's bank. Baxter assumes that the costs to support the service that Purchasers and Merchants jointly consume are distributed over their respective institutions. Based on this assumption, he concludes, "the geometry of aggregate supply is analogous to that of aggregate demand."

And as with joint demand, Baxter offers a depiction of the independent supply curves as well as the joint whole.

He then combines the separate geometric depictions of supply and demand into a figure that "depicts the resulting demand-supply equilibrium."

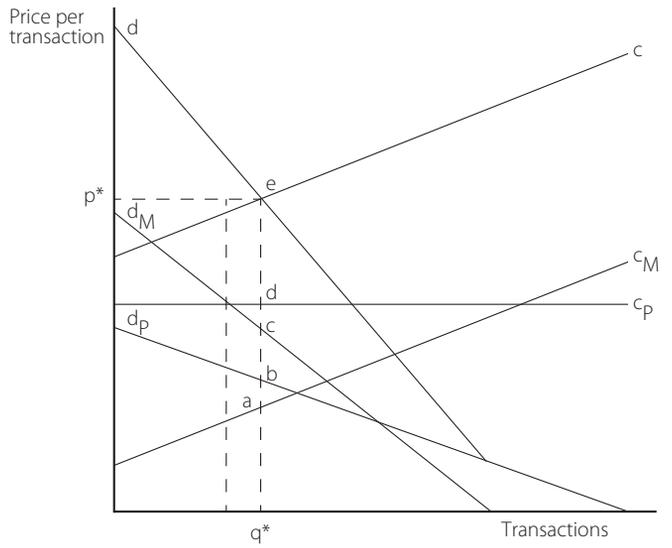


Figure 4: Merchant makes sales of amount S; M bank discounts q*c; merchant gets S – q*c; P bank collects S + q*b from purchaser; together banks retain (S + q*b) P + (–S + q*c) M = q*b + q*c = q*e; P bank remits S + q*b – q*d to M bank

This diagram sets up the paper's second critical insight:

*What is of critical importance is that the marginal cost q*d of the activities performed by the purchaser banks bears no necessary relation to the amount of revenue q*b forthcoming from the purchasers with whom those banks have contractual relationships.*

*Similarly, the costs q*a associated with the activities performed by merchant banks have no necessary relation to the amount of revenue q*c forthcoming from the merchants with whom they have contractual relationships.*

In other words, unless the banks on either side of the transaction are permitted to coordinate their joint supply decision through a side payment, from the side that collects too much to the side that collects too little, the four-party payment system will supply fewer transactions than is socially optimal.

Again, however, Baxter avoids presenting his conclusion in the language and form of formal economics. As Rochet & Tirole explain in a widely circulated draft of the famous paper noted above, those conclusions can be extracted from Baxter's analysis with a slight change in his notation. If benefits to purchasers and merchants are defined as marginal net benefits from the use of cards relative to other forms of payment, and if costs to purchasers and banks (or fees to their respective banks) are defined as marginal costs (or revenues), and if those banks are price-takers rather than price-setters, then as a payment system it will achieve the socially optimal level of output by setting the transfer payment between the two sides of the transaction at exactly the rate necessary to ensure that neither earns an economic profit.[7]

B) A BRIEF GUIDE TO BAXTER'S "BANK CREDIT CARDS AND THE INTERCHANGE FEE"

After theory comes history, at least in this article. Baxter takes the reader on an extended tour of the evolution of the check clearing system in the United States[8] and then returns the discussion to credit cards by way of merchant credit. "For centuries," Baxter observes, "merchants have extended short-term, interest–free credit to customers whose patronage is highly valued." And as Baxter explains, the rationale is quite intuitive. By making credit available to their best customers, merchants make it possible for consumer to (i) buy more on (ii) fewer visits and (iii) choose "higher-priced items" than they might otherwise.

Baxter then points to a shift from merchant credit to third-party credit following World War II. He posits the existence of a "frequent traveler" with "high income and high time costs" who would have access to local merchant-supplied credit at home but not on the road. This hypothetical "frequent traveler" would have poor payment options available to him—cash, traveler's checks, and personal checks. Cash, from a consumer perspective, carries a significant risk of loss. Traveler's

checks come with "high time costs." Checks require the presentation of "identification at a moment when time costs [are] greatest" (i.e., the moment of purchase) and "not infrequently" involve "humiliat[ion]" with the effort to confirm identity at the point of sale.

According to Baxter, card-based payment systems arose to meet this demand. Non-banks such Diner's Club and, later, American Express offered three-party systems. Such systems, of course, did not need an interchange mechanism. One firm both signed merchants to accept cards and issued cards to consumers. Four-party bank systems came later. Three-party bank systems that had evolved in specific geographies became four-party to achieve ubiquity that "by reason of our geographically restrictive banking laws, could not be obtained by any single banking enterprise."

Having laid out the four-party model earlier, Baxter then delivers the rhetorical coup de grâce on the need for an interchange mechanism:

[M]ultibank organizations were from their inception four-party systems having the peculiar economic characteristic previously described. Given the distribution of charges between P and M that would achieve equilibrium in their demands, it was overwhelmingly improbable that the revenue stream from M to M bank or from P to P bank would equal the costs of the subset of activities that a particular bank was required by the technology of the payment system to perform; thus some redistribution of those revenues between M bank and P bank was likely to be necessary for the payment system to compete effectively with alternative mechanisms.

Although the article--or, at least, the section--could end there, it does not. Baxter proceeds to answer three discrete questions: (1) whether individual bank negotiations might take the place of centrally set interchange; (2) whether interchange fees should be set at 0 (as in the checking system); and (3) whether interchange rates are currently set at the socially optimal level. The article does not, however, attempt to motivate the discussion, and it seems, at least without context, a bit forced.

Context is, however, available. These questions flow directly from the litigation that served as the inspiration for Baxter's paper, NaBanco v. Visa U.S.A., Inc.[9] NaBanco argued (1) that individual negotiations between counterparties to specific transactions could take the place of centrally set interchange; (2) that the court should simply set interchange at 0, effectively allowing acquirers to keep the entirety of what they collect from merchants; and (3) that interchange had been set "too high."

The article's answers to these questions are not entirely satisfying. The article marches through them as if it were following an indisputable chain of logic. But the explanations are not entirely persuasive. The problem is largely rhetorical. After asserting that interchange is necessary for a four-party payment card system, Baxter's writing becomes significantly more conditional. Key sentences throughout the discussion use words like "can," "could," and "possible." And as in the theoretical section that opens the piece, Baxter eschews external references.[10] In at least this respect, the court's discussion of these points is more satisfying. The court rejects NaBanco's efforts to replace interchange with individual bi-lateral negotiations by observing that the transaction costs in such a system would be "high and stultifying."[11] The court similarly dismisses the claim that the Sherman Act requires interchange fees to be set to $0. Using more or less the same verbal formulation that Baxter's article uses to introduce interchange, the court explains that nothing in the system "suggests, much less guarantees" that revenue streams on either side of the system will be sufficient to cover the costs unique to that side of the platform.[12] The court also has little patience for the argument that Visa arrived at its interchange rate through a flawed process. As the court explains, although the process through which Visa set interchange may not have been perfect, it "was and is careful, consistent, and within the bounds of sound business judgment."[13]

II. FINAL THOUGHTS

Baxter's paper is the first scholarly paper to discuss a tool that helped propel the rise of electronic payments around the world and that has been the subject of nearly constant legal and regulatory scrutiny since its introduction nearly forty years ago. With the passage of time, it has become difficult to separate Baxter's contribution from those who helped to formalize and extend his work.[14] But even if lawyers and economists interested in interchange and payment card networks must look beyond Baxter for answers to their questions, his article remains, even after the passage of time, the best place to start.

1 *See* Jean-Charles Rochet & Jean Tirole, *Cooperation Among Competitors: Some Economics of Payment Card Associations*, 33 Rand J. Econ. 549, 564 (2002).

2 Dennis W. Carlton, *Externalities in Payment Card Networks: Theories and Evidence* 126, *in* Kansas City Federal Reserve, The Changing Retail Landscape: What Role For Central Banks? (2010).

3 *See, e.g.*, Benjamin Klein et al., *Vertical Integration, Appropriable Rents, and the Competitive Process*, 21 J. L. Econ. 297 (1978).

4 The trepidation arises from the quasi-religious devotion and disdain that Baxter continues to inspire in fans and critics. *Compare* Richard A. Posner, *Introduction to Baxter Symposium*, 51 Stan. L. Rev. 1007 (1999) (recalling his impression upon meeting Baxter in 1967 when interviewing for a junior faculty post at Stanford—"I was instantly, immensely, and permanently impressed by the power of his mind and the clarity of his expression") with Lloyd Constantine, Testimony Before The Antitrust Modernization Commission (2005), *available at* http://govinfo.library.unt.edu/amc/commission_hearings/pdf/Constantine.pdf (complaining that Baxter, who as head of the Antitrust Division had reportedly defied President Reagan in pursuing the case against AT&T, had "prophe[sied]" and sought to eliminate "federal antitrust enforcement").

5 *See* Gary S. Becker, Economic Theory 85 (1971) (discussing "the Marshall-Pigou tradition" and formally describing models with externalities running between competing firms).

6 *See* John Nash, *Two-Person Cooperative Games*, 21 Econometrica 128 (1953).

7 *See* Jean-Charles Rochet & Jean Tirole, *Cooperation Among Competitors: The Economics of Payment Card Associations* 4-5 (May 16, 2000), available at http://www.wcas.northwestern.edu/csio/Conferences/CSIO-IDEI-2000/tirole.pdf. The working draft also explicitly credits Baxter for observing that in a four party system "there is no reason why both banks should break even on the transaction." *Id*. at 4.

8 As discussed above, this essay is going to skip Baxter's discussion of the check system. That section of the article is well footnoted and generally straightforward. Moreover, although some critics harbor objections to some elements of Baxter's history, see, e.g., Alan S. Frankel, *Monopoly and Competition and Exchange of Money*, 66 Antitrust L.J. 313 (1998), Baxter's discussion of the evolution of the check system is generally regarded as authoritative.

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

10 Two very short sections follow Baxter's discussion of interchange in the credit card systems. The first is labeled "Modern Developments," and the second is simply "Conclusion." The section devoted to "Modern Developments" offers some predictions about debit cards that, at least in the wake of the cases challenging Visa's and MasterCard's respective honor all cards rules proved quite prescient—on page 585, Baxter notes, "It seems likely . . . that the two payment vehicles [debit and credit] will have to be differentiated and subjected to different patterns of distributing charges between merchants and card holders and, in all probability, to different interchange fees." The Conclusion contains very truncated discussions of two key legal issues—(1) whether arrangements setting interchange should be viewed as price fixing; and (2) whether further cooperation among the banks that make up the card networks should be condoned. According to Baxter, the answer to both questions is no.

11 *Nabanco*, 596 F. Supp at 1261.

12 *Id*. at 1260.

13 *Id*. at 1262.

14 Since Baxter's piece was published, it has been cited by 260 papers and legal opinions in English and other readily scannable languages. A full 170 of those citing works also cite the work of David Evans, Jean-Charles Rochet, Richard Schmalensee or Jean Tirole.

BANK INTERCHANGE OF TRANSACTIONAL PAPER: LEGAL AND ECONOMIC PERSPECTIVES

William F. Baxter*

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

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

I. THE THEORETICAL VIEWPOINT

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

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

---

* This article was originally published in volume XXVI of the Journal of Law & Economics. At the time, William Baxter was Assistant Attorney General of the Antitrust Division in the United States Department of Justice. In his introduction, Baxter wrote, "This paper was written while I was Professor of Law at Stanford University and revised thereafter. The views expressed here are my own and are not official policy statements of the Antitrust Division or the Justice Department. I thank J. Anthony Chavez and Greg Sidak for their helpful research assistance and suggestions." The article is reprinted with the permission of the University of Chicago Press.

A) THE DEMAND FOR TRANSACTIONAL PAPER

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

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

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

Demand for a private good depends on each person's evaluation of the good's marginal utility and can be described by a function indicating the amount of product the person is willing to buy at a given price. Each consumer's evaluation of the marginal utility of a private good is usually independent of other consumers' evaluations, and so aggregate demand at any price level is the sum of the individual demands at that price. For example, if the prevailing price of shoes is $30 a pair, consumer Jones will buy one, and then another, and then another pair of shoes until the marginal value he attaches to the next pair (which he does not buy) falls below $30. The same is true for consumer Smith, although there is no reason to expect that at any particular price each will demand the same number of pairs, because there is no particular reason to suppose that the marginal value that Jones attaches to the third or fifth or eighth pair of shoes is the same as the marginal value that Smith attaches. Because the evaluations of the marginal value of shoes by Jones and Smith are independent of one another, the aggregate demand of Jones and Smith for shoes at $30 a pair is simply the sum of their individual demands at that price.

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

This identity in the type of transactional service used by the merchant and purchaser in a given exchange introduces a constraint not normally found in markets for private goods and reflects the interdependence in the marginal valuations between merchants and purchasers. Because the mechanics of transactional services require the acceptance of a particular payment mechanism by **both** the merchant and the purchaser

to effect any given purchase, the marginal valuation of a transactional service by one party to the purchase is contingent on the acceptability of this form of service by the other party. On the one hand, given that particular payment mechanism is acceptable to the other party, marginal valuation is determined in the usual manner for private goods. On the other hand, if the payment mechanism in question is unacceptable to the other party for whatever reason, the marginal valuation by the first party is zero regardless of the magnitude of its value when the mechanism is acceptable. The contingent nature of these marginal valuations of transactional services by merchants and purchasers, and hence the contingent nature of the individual demands for these services, destroys the independence necessary to permit the calculation of aggregate demand by summing the individual demands horizontally and largely renders intractable the economics of transactional paper in this particular description of the market.

Perhaps the most intuitively appealing way to resolve the difficulties posed by this market model is to redefine what we mean as one unit of the product consumed. Rather than considering the demands of *P* and *M* as demands for separate products, define one unit of product to consist of the bundle of transactional services that banks must supply jointly to *P* and *M* in order to facilitate the execution of one exchange of goods or services between *P* and *M*. Under this interpretation, the supply price of the product is the sum of the individual charges to *P* and to *M*. Furthermore, the demand for that product is a joint demand of *P* and of *M*: in combination they must make a payment of that magnitude to the banks to induce the necessary supply, but independently neither *P* nor *M* necessarily confronts any particular price as one he must pay in order to have his demand fulfilled.[4] This model preserves the excludability property of transactional services.

Figure 1 illustrates the derivation of aggregate demand for transactional services of a given type in a single-merchant, single-purchaser economy. The quantity axis is calibrated in units which represent the bundle of services that must be provided by banks to both *P* and *M* in order to facilitate one exchange. The vertical axis gives the reservation prices of the two traders for various levels of consumption of the transactional services. Line d_M represents the demand schedule of *M* for such complete units of transactional service on the assumption that *P*—*M*'s customer—is willing to use this particular service but unwilling to make any contributory payment for the units when purchased from the bank.

Line d_p represents the demand schedule of *P*, based on the assumption that *M* is unwilling to make any contributory payment for those services. Given the information shown in line d_M and line d_p, the aggregate demand schedule of *M* and *P* for these units of transactional services is line *d'*, which is obtained by summing vertically the separate demand schedules of *M* and *P*. In other words, the schedule *d'* is constructed so that if any vertical line is drawn through the figure, the distance v_1v_4 equals the sum of distances v_1v_2 and v_1v_3.

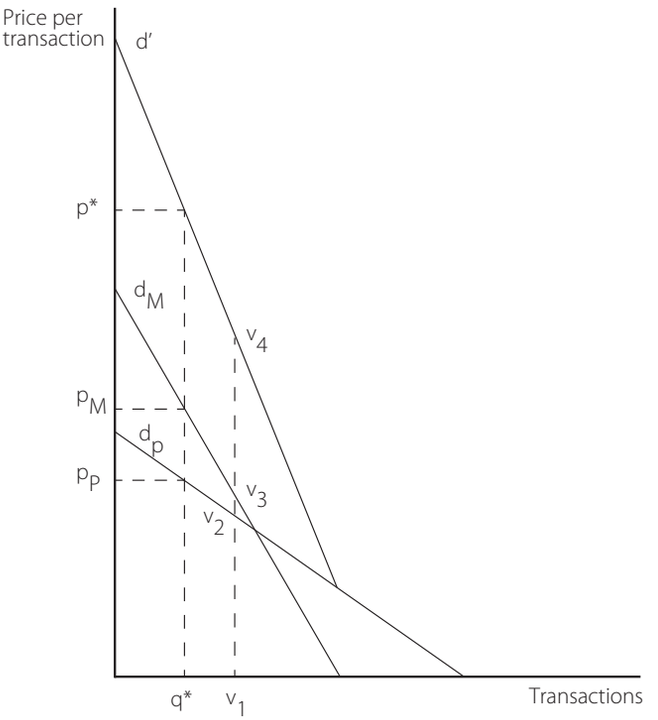


Figure 1

Figure 1 should be interpreted as follows: if the price per complete transaction—that is, the total revenue banks will demand to provide the services necessary to facilitate one exchange between *M* and *P*—is *p**, then the quantity of transactions that *M* and *P* should demand is *q**, the quantity indicated by a vertical line dropped from the intersection of *p** and *d'*. I say "should" rather than "will" be demanded because, although *q** is the quantity of transactions that maximizes the aggregate benefits of *M* and *P*, a certain amount of coordination is prerequisite to *M* and *P*'s arriving at that outcome. Specifically, this favorable outcome will result only if the aggregate price *p** is apportioned between *M* and *P* in the proportions represented by the height of their respective demand curves at output level *q**. That is, for each transaction, *P* must find a way to make some payment p_p to the banks, and *M* must find a way to

make some payment p_M to the banks; when p_P and p_M are summed they will, by construction in Figure 1, equal p^*, the price that the banks demand for providing those services. If there are no bargaining costs-that is, if *P* and *M* have perfect information and neither persists in strategic bluffing to reduce his own costs at the expense of the other-they would bargain to this particular outcome. On the other hand, if either *P* or *M* strategically insists on paying less, then, because the other can be induced to pay no more at so high a level of transaction services, both *P* and *M* will be harmed, for the sum of their contributions will be less than p^*; thus the banks will decline to provide services that *M* and *P* together value at p^*.

One must resist any impulse to say that *M* is paying too much and *P* too little in the circumstances depicted by Figure 1. Given that the banks will insist on receiving revenues per transaction in the amount p^*, and given that *P* is unwilling to pay more than p_P per transaction at output level q^* for the very good reason that he does not value the service any more highly, *M* can only worsen his position by declining to make a payment per transaction in the amount pm. For it is inescapable that *M* and *P* must agree on some specific number of transactions to be effected by the payment mechanism in question. And if that number is to be q^*, then in our hypothetical case depicted in Figure 1 agreement can only be reached if *M* is willing to pay the preponderant share of the price p^*. In the region q^*, *M* values the marginal transaction more highly than does *P*, and *M* pays accordingly.

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

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

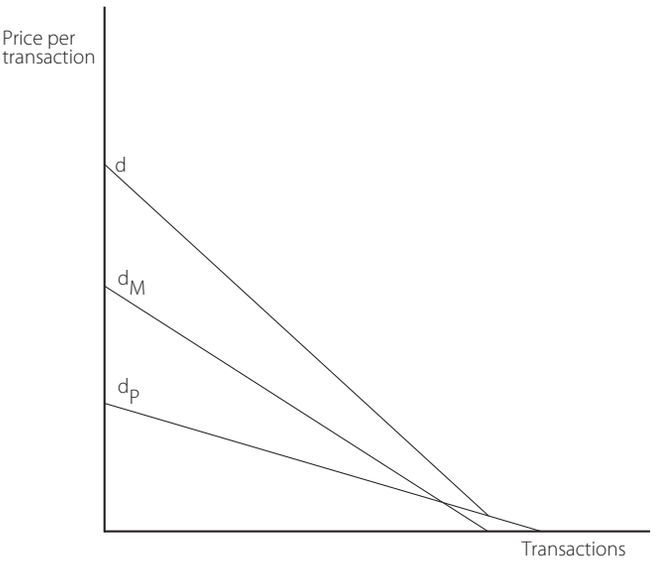


Figure 2

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

B) THE SUPPLY OF TRANSACTIONAL PAPER

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

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

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

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

Correspondingly, the geometry of aggregate supply is analogous to that of aggregate demand. It is conventional to think of the supply curve for an industry as being constituted by the horizontal aggregation of the supply curves of the individual firms. But because the costs incurred by the banks are joint, when *P* bank participates on behalf of purchasers and *M* bank participates on behalf of merchants, the costs of the two firms must be aggregated vertically, not horizontally, in order to obtain an analytically useful representation of the full marginal cost per transaction and hence of the number of purchaser-merchant exchanges that banks will facilitate at any particular price level for transactional services.

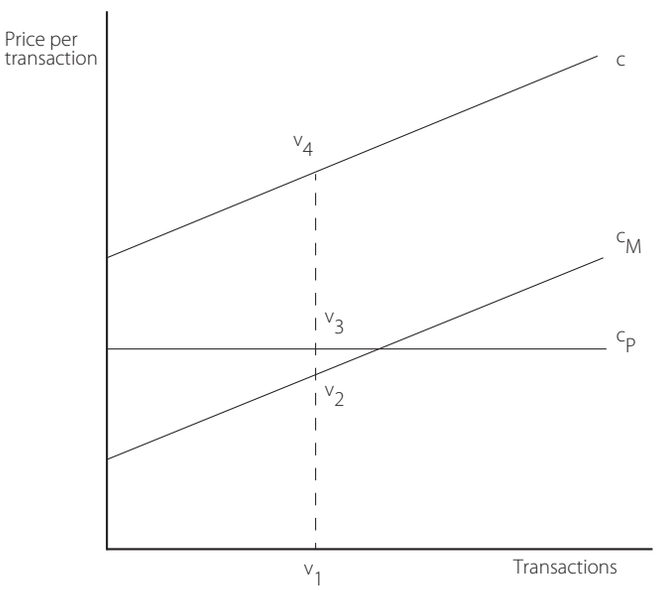


Figure 3

Figure 3 depicts possible marginal cost curves c_M for *M* bank, and c_P for *P* bank, together with their vertical aggregation *c*, which corresponds to the total marginal cost per exchange facilitated by the two participating banks. As before, the technique of vertical aggregation is such that, given any vertical line drawn through the curves, the distance v_1v_4 equals the sum of the distances $v_1v_2 + v_1v_3$.

Somewhat arbitrarily, I have drawn Figure 3 in a way that suggests that *P* bank's costs exhibit constant returns to scale whereas *M* bank's costs exhibit decreasing returns to scale, but nothing in the analysis turns on those particular assumptions.[7] Figure 3 also could be thought of as depicting industry supply, if one views c_P as a traditional horizontal summation of the marginal cost curves of all purchaser banks, and c_M as the traditional horizontal summation of marginal cost curves of all merchant banks. But in this interpretation, too, the vertical summation *c* of those two sets of costs depicts the industry supply curve, for with respect to each transaction, revenue equal to *c* must be forthcoming in order to cover all industry marginal costs

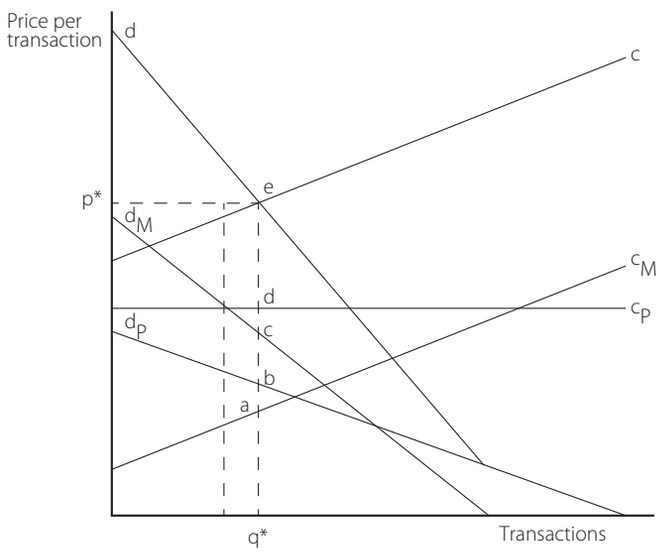


Figure 4—Merchant makes sales of amount *S*; *M* bank discounts *q*c*; merchant gets *S* – *q*c*; *P* bank collects *S* + *q*b* from purchaser; together banks retain (*S* + *q*b*) *P* + (–*S* + *q**c) *M* = *q*b* + *q*c* = *q*e*; *P* bank remits *S* + *q*b* – *q*d* to *M* bank. At close,

P's position	*–S*	*–q*b*			
P bank position	*+S*	*+q*b*	*–S*	*–q*b*	*+q*d*
M bank position	*–S*	*–q*c*	*±S*	*+q*b*	*–q*d*
M's bank position	*+S*	*–q*c*			
Totals down	0	0	0	0	0
Totals across:					
P bank	*+q*d* = cost				
M bank	*q*c + q*b – q*d = q*a* = cost				
"Interchange fee"	(*q*d – q*b*) = (*q*c – q*a*)				

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

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

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

It is true, of course, that a side payment of ac per facilitated exchange from *M* bank to *P* bank is not the only conceivable institutional adjustment, but it appears to be by far the simplest and the least expensive.[9] Since any redistribution mechanism will itself involve a transaction cost which will serve to raise *C*, the

mechanism that minimizes transaction costs is in the interest of all the parties. Since remittance of funds in some amount from *P* bank to *M* bank is an inescapable feature of any payment mechanism of the type under consideration, adjustment of the magnitude of that remittance to achieve the equilibration of costs and revenue clearly appears to be the preferred mechanism.

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

One important assumption underlies the preceding paragraph: banks participating in the payment system are behaving competitively and charging prices to *P* and *M* corresponding to the bank's marginal costs and, in equilibrium, to their average total costs including the opportunity costs of invested capital. There are two quite distinct reasons why this assumption may not hold in any particular real world context. First, through collusion the banks might have acquired enough market power to be able to charge both purchasers and merchants prices that exceed the banks' cost.[10] I explore the implications of collective action among banks more fully, later in this paper.[11] For the present, I note only that the problem of cartel profit maximization will be complicated by the fact that, in order to maintain an equilibrium number of transactions, the cartel must increase prices each to merchants and to purchasers in amounts dictated by the slope of their demand curves—amounts that, in all probability, are equal neither in absolute magnitude nor in percentage markup over the competitive price. Hence cartelization of the industry would be comparatively difficult.[12]

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

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

II. THE HISTORY OF FOUR-PARTY TRANSACTION VEHICLES

Over the last 150 years, three distinct categories of four-party cashless payment systems have evolved. The check and the bank credit card are heavily used today to facilitate exchanges, and the debit card is increasingly being promoted. This section presents a brief history of the commercial environment in which each of these

developed in conjunction with each of them. By use of the economic theory developed in Section I, it is possible to uncover previously unrecognized forces in the evolution of these payment systems.

A) THE PRACTICE OF PAYING CHECKS "AT PAR"

In the early 1800s the two principal means of payment in commercial transactions were (i) bank notes issued by state banks and (ii) drafts. These two media can be thought of as corresponding to (i) currency and (ii) checks today. Although checks had an early origin,[13] they did not become common until after the Revolutionary War.[14] In the years between the demise of the Second Bank of the United States and the Civil War, checks were commonly used as a means of paying local bills only in the nation's commercial centers.[15] City banks encouraged the use of deposit currency because inferior country bank notes of uncertain value tended to drive the sounder city bank notes out of circulation.[16] For the most part, the attempts of the city banks to prevent the discounting of these notes were unsuccessful.[17] During this time, transportation outside the nation's commercial centers was slow, expensive, and often dangerous. Only infrequently did either goods or people travel very far. Markets were predominantly local, and goods consumed in any geographic area usually had been produced there.

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

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

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

In 1864 Congress passed the National Bank Act,[22] reinstituting the rivalry between state and national banking systems that had existed during the nation's first half century. Federal taxes were levied on bank notes issued by state banks in an endeavor to drive the notes, and perhaps the banks, out of existence.[23] Although the 1864 Act required that national banks maintain reserve deposits, it permitted a large fraction of those reserves to be held as deposits in designated "reserve banks" in various major cities; and, because drafts could be issued against these reserves, the national banking system became instrumental in the payments system.[24]

The era was one of rapid technological change in both transportation and communications. The railroads, waterways, and post roads expanded rapidly, frequently under the spur of government subsidies, and the telegraph was invented and deployed. These changes tend to explain the increase in use of transactional paper relative to currency, but it is less clear why the use of checks relative to drafts also increased very rapidly during this period.[25] When a check was used to pay a distant payee, *P*, having a positive balance with *P* bank,

sent the instrument (usually by mail) to *M*, who presented it to *M* bank for collection. Then *M* bank accepted the instrument for collection and might or might not credit *M*'s account with *M* bank for the amount of the check before collection had been achieved.[26] The instrument was started by *M* bank on what was often a circuitous journey from one bank to another until through some series of correspondent relationships it arrived at *P* bank.[27] The check was accepted by *P* bank and debited against *P*'s account. At this point *P* bank again faced the problem of making payment to *M* bank, just as when drafts were used. Again, its costly alternatives were the actual transport of currency or the maintenance of geographically dispersed balances against which a draft in favor of *M* bank could now be issued.

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

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

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

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

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

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

For some or all of these reasons it seems to have been necessary for the industry to redistribute the direct costs of the checking system away from *P* and toward *M* so that the market for transactional paper could equilibrate. That need may itself best explain the relatively sudden displacement of the draft by the check. A new and less familiar instrument, the check was accompanied by

fewer customs and fixed expectations than the more familiar draft. And the check, although very similar to the draft in most respects, passed through the hands of the four parties in a different sequence, a sequence that tended to enhance monopsonistic position of *P* bank as a buyer of paper.

As Figure 4 demonstrates, if the level of total banking costs (and therefore the values of p^* and q^*) changed significantly, then no change in the aggregate demand curve of *P* and *M* would be necessary to change the relative magnitudes of their individual demand levels for use of a payment system. It is well established that from the Civil War to the end of the nineteenth century p^* fell by a considerable amount and q^* increased enormously.[30]

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

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

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

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

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

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

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

Notwithstanding the advantages of uniform (perhaps uniformly zero) exchange charges, a very large number of banks strenuously resisted remitting at par. The banks that continued to charge exchange into the twentieth century were, almost without exception, small banks in isolated agricultural communities. For the banks that adhered to this practice, revenue in 1964 from exchange charges constituted about 10 percent of total current

operating revenue, and the percentage was higher for the smaller institutions among the group.[35] It seems likely that in the late 1800s and early 1900s, when the nonpar controversy was at its height, this form of income was even more important to the small country bank.[36]

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

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

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

After the monetary panic of 1907, a national monetary commission was appointed to study the American banking system.[37] Its report led to the passage of the Federal Reserve Act in 1913.[38] This legislation, its subsequent amendments, and the practices and rules of the Federal Reserve Board, which the legislation created, eventually tipped the balance in favor of par clearance in the United States. It was not obvious from the initial legislation that this outcome would result, nor is there any reason to believe that the practice of nonpar banking particularly concerned either the National Monetary Commission or the Congress of 1913.[39] The key provisions of the Federal Reserve Act were sections 13 and 16. Section 13 initially read, in part:

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

Section 16 read, in part:

Nothing herein contained shall be construed as prohibiting a member bank from charging its actual expense incurred in collecting and remitting funds, or for exchange sold to its patrons. The Federal Reserve Board shall, by rule, fix the charges to be collected by the member banks from its patrons

whose checks are cleared through the Federal reserve bank and the charge which may be im- posed for the service of clearing or collection rendered by the Federal reserve bank...[41]

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

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

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

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

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

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

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

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

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

In 1918 the Fed succeeded also in obtaining from the attorney general an opinion that in effect prohibited precisely what the Hardwick Amendment seems, at first glance, to have permitted. Focusing on the last few words in the Amendment, the attorney general ruled that the federal reserve banks were prohibited by law from paying, even in the sense of passing on, exchange charges in the course of the clearance process.[52]

Since, in the period under discussion, the system would not accept items drawn on nonmember banks not on the par list, the clause, even thus interpreted, would appear to have been inconsequential. But the Fed made it of consequence in 1919, adding substantially to the number of banks on the par list by introducing a new coercive device.

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

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

[The] proviso in Section 13... has been constructed by the Attorney General... as meaning that a Federal reserve bank cannot legally pay any fee to a member or non-member bank for the collection and remittance of a check. It follows, therefore, that if the Federal reserve banks are to give the service required of them under the provisions of Section 13 they must, in cases where banks refuse to remit for their checks at par, use some other means of collection, no matter how expensive.

The action of the various Federal reserve banks in extending their par lists has met with the cordial approval the Federal Reserve Board, which holds the view that under the terms of existing law the Federal reserve banks must use every effort to collect all bank checks received from member banks at par. Several of the Federal reserve banks are now able to collect on all points on their respective districts at par, and new additions to the other par lists are being made every day. The board sees no objection to one bank charging another bank or a firm or

individual the full amount provided in Section 13 of the Federal Reserve (10 cents per $100) and has not undertaken to modify these charges, but the Act expressly provides that no such charge shall be made against the Federal reserve banks.[56]

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

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

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

B) BANK CREDIT CARDS AND THE INTERCHANGE FEE

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

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

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

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

Open book credit well served the parties affected while travel outside one's home community was relatively infrequent. After World War II, the frequent traveler was likely to have a high income and high time costs and therefore to have been extended open book credit in his own community; but away from home he could not readily be identified at the point of sale. He could carry large amounts of cash, but the risk of loss was substantial.

Traveler's checks were an alternative, but they involved high time costs because they required the traveler, first, to visit the bank before departing and, second, to predict with reasonable accuracy how much money would be needed during the trip or to make another journey to the bank on return to redeem the excess checks, or to leave funds tied up on a non-interest-bearing certificate until a later time when the traveler's checks might be used. A second alternative—attempting to cash personal checks at one's destination—involved tediously presenting identification at a moment when time costs were likely to be greatest; not infrequently, the attempt was humiliatingly unsuccessful. From the standpoint of the merchant located at the traveler's destination, the situation was also unsatisfactory. If the merchant could easily identify the traveler as a creditworthy consumer with high time costs, he would be only too happy to extend to the traveler the same credit facilities extended to comparable local customers.

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

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

Because the T&E card was a three-party instrument rather than a four-party instrument, the feature of jointness was present on the demand side but not on the supply side. Again, there was one particular distribution of costs between the merchants and the card holders that would bring their demands for the transactional service into equilibrium. But the card-issuing organization was a single enterprise; periodic adjustment was within its control, and there was no problem of coordinating two enterprises to determine how to distribute charges between card holders and merchants.

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

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

Moreover, as nominal interest rates began to rise by the late 1960s, interest costs became a larger fraction of the total cost of extending consumer credit. The comparative advantage of banks and other financial institutions over all but the very largest of the retail chains became ever more decisive as interest costs predominated in the total cost of performing the retail credit function. Finally, there were scale economies from consolidating one consumer's transaction with a number of merchants into a single statement, a single billing, and a single remittance.

All these factors favored substituting bank-card systems for the traditional merchant function of extending retail credit.

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

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

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

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

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

Imagine a card system composed of ten *P* banks that act only as purchaser banks and ten *M* banks that act only as merchant banks. Assume that each *P* bank receives from each *M* bank 1 percent of the aggregate paper flow of the system and has 10 percent of the aggregate card-holder base. Assume, finally, that the optimum transfer fee is 1 percent of the face value of the paper and that this fee amounts to $0.30 per item. Although it is subversive of the system as a whole to demand a higher fee, each individual *P* bank faces a strong temptation to do so—let us assume a 10 percent increase in the transfer fee to 1.1 percent, or $0.33. Any individual *P* bank that so behaves, provided that it is unique in demanding an excessive fee, will increase its fee revenues by about

10 percent but will increase the effective costs confronted by each *M* bank only by 1 percent. Even assuming that the *M* banks immediately pass on this cost differential, the merchant discount would be increased by 1 percent on the paper of all *P* banks, for it is not feasible for the *M* banks to discriminate against paper en route to that particular *P* bank without creating, on the part of all the merchants, an incentive to refuse to honor cards issued by that *P* bank; moreover, any endeavor by all merchants selectively to refuse cards issued by a particular *P* bank (at least outside the context of an on-line electronic system) would substantially increase the transaction costs of all merchants and of all card holders. The utility of the system to all participants would diminish, as would the system's viability in competition with other payment systems.

Similar, although perhaps less immediately dramatic, consequences would follow if either the set of *M* banks or the set of merchants chose to absorb the percent cost increase that flows from *P* bank's 10 percent increase in the transfer fee. Some might drop out of the system entirely because of economic losses; others would alter their behavior in less drastic ways to shift from using the card system to using some other payment systems. These adverse consequences would eventually reduce the transaction volume of the individual *P* bank that raised the transfer fee, but the adverse effect would be spread across all *P* banks. The one *P* bank would realize 100 percent of the revenue gains from its fee increase but would bear only 10 percent of the adverse consequences. More generally, in a card system involving *x* number of *P* banks, any one bank can exploit the monopsonistic position it enjoys over its own paper and can realize 100 percent of the revenue gains while suffering only a fraction of the adverse consequences, that fraction being 1/*x*. Accordingly, it is essential that the participants in a four-party payment system collectively adopt some internal mechanism that prevents individual exploitation of the monopsony power endemic to such systems.

As discussed earlier, banks were prevented from exploiting their monopsonistic power in the checking system initially by collective agreements among clearinghouse members and later by the Fed's coercive tactics. But the problem was resolved for the checking system without explicit recognition of the problem's characteristics, without any inquiry into the costs of the system, at the apparently arbitrary transfer fee of zero, and largely by government coercion rather than agreement. These all make it unlikely that the resolution was optimum when first made, even less likely that the resolution could have continued to be optimum after the enormous changes in check-processing technology. Compared to the checking system, the bank credit card system has evolved so far under less government intervention with respect to the transfer fee. Perhaps for that reason, perhaps also because there are many institutions for which items transmitted in their capacity as *M* bank are unequal to items received in their capacity as *P* bank, behavioral characteristics of those payment systems more closely correspond with the behavior implied by the theoretical considerations discussed in Section I.

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

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

Clearly, *P* bank bears the cost of this extended period of float, but the incidence of the corresponding benefit on demand is ambiguous. In comparison with use of a currency or a check method of payment, *P* is the beneficiary, and his demand for the bank-card system should increase. On the other hand, to the extent that the bank-card system is being used by *P* and *M* in lieu of open-book credit, it is *M* whose float costs have been reduced, and his demand should be enhanced.

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

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

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

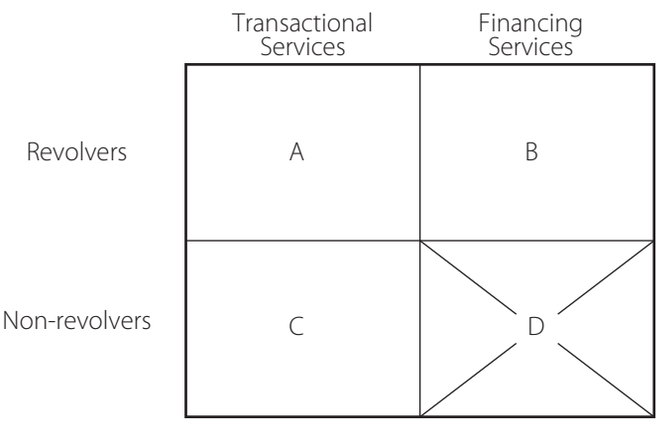
With exceptions to be discussed later, no charge has been imposed on the card holder. In this context, too, each card-issuing bank is free to negotiate such arrangements as it wishes with its card holders. Before 1980 only a few card-issuing banks had imposed either transaction fees or periodic "membership" fees on their card holders; in the overwhelming preponderance of instances, banks have been willing to play the role of *P* bank as a competitive gambit to attract the individual demand deposits of its card holder. Until recent regulatory reform permitted banks to pay interest on demand deposits, the value to the card-issuing bank of attracting incremental individual demand deposits on which no interest was or could be paid was a sufficient inducement, at least when coupled with the interchange fee received from the merchant bank, to compensate *P* bank. Thus, although revealed demand plainly exists on the merchant side, it is less clear on the card holder side.

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

Thus the card-issuing bank can be viewed as engaged in two different businesses. It sells a transaction service involving valuable float to those "nonrevolvers" who choose to pay their statement in full at the end of each billing cycle. It also sells a combination transaction service and consumer finance service to those who use their bank cards as an extended credit mechanism.

Because certain activities essential to providing the payment service—receipt of interchange items, posting to individual card holder accounts, billing, collection, posting of credits, bearing the risk of default, etc.—must be performed with respect to revolvers as well as nonrevolvers, complex accounting allocation problems arise.

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



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

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

Alternatively, one could view the industry as being engaged in two businesses but, rather than linking cell A with cell C and defining cell B to be the separate business, this second view links cell A with cell B and defines cell C to be a separate business. This view defines the two businesses with reference to card holder payment practices. One business consists of providing transaction and financing services to revolvers; another consists of providing transaction services to nonrevolvers. The implied accounting allocation problem is to allocate each category of banks' activity costs either to revolvers as a group or to nonrevolvers as a group. Under this view, the cost allocation problem is to associate some fraction of total bookkeeping costs and total fraud and default costs with habitual revolvers and the remaining fraction with habitual nonrevolvers.

For habitual revolvers, there are three possible revenue sources: periodic interest charges on outstanding balances, the merchant discount, and other card holder charges such as membership or per dollar fees. For nonrevolvers, only the two latter revenue sources are available.

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

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

C) MODERN DEVELOPMENTS

Several events since 1980 require significant adjustments by the bank- card industry. Among the most important are the changes introduced by the Depository Institutions Deregulation and Monetary Control Act of 1980.[64] This legislation, and the regulations that implement it, require the Fed to impose cost-based fees on banking institutions to which it renders services, including check-clearing and collection service; authorize the Federal Home Loan Bank Board to render clearing and collection services, again on a cost-based fee basis, to savings and loan institutions (S&Ls); authorize a significantly broadened scope of activities by S&Ls, including nonbusiness demand deposits (NOW accounts), broadened lending authority, and credit card services; and authorize both banks and S&Ls to pay interest on demand deposits.

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

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

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

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

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

Simultaneously, however, the supply curve for credit card transactions will also be shifting to the right because of the high cost of funds. Not only the height of these functions but also their shapes over the relevant range will undoubtedly change in ways we do not yet know. As I emphasized in Section I, the shifting cost function under consideration cannot usefully be viewed as reflecting the cost of dealing with card holders; it reflects the joint cost of providing transaction services to both card holders and merchants. Nevertheless, substantially all of the recent price changes are in the charges imposed on card holders rather than in the merchant discount.

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

In addition to the present perturbations in the industry, the "debit card" is for the first time being distributed widely. Apparently many institutions in the industry believe that the debit card and the credit card can be combined and embodied in a single set of plastic cards. Transactions using the cards would be subject to the same merchant discount and the same interchange fee notwithstanding that the card-issuing bank would handle the two types of transactions quite differently. This outcome seems most unlikely unless the contractual terms that have traditionally accompanied the credit card are materially altered. From the standpoint of the card-issuing bank, debit card transactions will be substantially cheaper than credit card transactions, for debit card transactions will not be authorized unless they are for amounts less than the card holder's deposit balance, in which case the default risks are relatively low. Moreover, since the transaction amount is immediately debited against the card holder's deposit balance, the float costs of the debit card are substantially less. These considerations alone seem to dictate quite a different distribution of fees between card holder and merchant and a different interchange fee, as well. In addition to these cost factors, demand factors suggest a similar conclusion. From the card holder's standpoint, the debit card is less attractive than the credit card. The float costs that the bank saves when a debit card is used are precisely the float benefits that the card holder forgoes when he uses a debit card. One would expect therefore that any card holder entitled to use a credit card will always use it rather than a debit card. It follows that the only frequent users of debit cards will be people whose incomes and other indicators of creditworthiness do not enable them to obtain and use credit cards.

The characteristics that distinguish credit card users from debit card users will substantially affect the demand curve of merchants for transactions with these two different types of card holders. The holder of a credit card will continue to be identified as a customer for whose patronage the merchant wishes to compete by extending a free float period; but that will not be true of the holder of a debit card, and one would expect merchants to be unwilling to accept discounts on debit card paper as large as the discounts traditionally accepted on credit card paper.

It seems likely, therefore, that the two payment vehicles will have to be differentiated and subjected to different patterns of distributing charges between merchants and card holders and, in all probability, to different interchange fees. Hence I believe that card-issuing institutions will be engaged in not one but two different learning processes in the period immediately ahead; and both processes will be retarded if these institutions are reluctant to recognize the sharply different cost and demand characteristics of the two payment vehicles.

III. CONCLUSION

Four-party payment vehicles such as the check, the credit card, and the debit card are characterized by joint costs and also by interdependent demand on the part of their users, which, despite the antiquity of such mechanisms, neither the economic literature nor the institutions that provide their services have fully recognized. Those characteristics, in my judgment, were an important contributing cause to the controversy over "clearance at par" that troubled the banking industry for more than half a century and was quieted at last only by means of federal coercion and subsidy. A repetition of the same basic controversy in the context of new payment mechanisms—credit cards and debit cards—is likely to occur in the next few years. Because of sharp cost and demand changes attributable to legislative amendments, because of the effect of inflation on nominal interest rates, and because of governmental responses to inflation that have taken the form of restrictive monetary policies that increase the real interest rates on short-term obligations, those years are likely to be characterized by disequilibrium, confusion, and controversy. In such a period, reliance on governmental intervention to reduce uncertainty is likely to appeal to at least some of the disputants. Such intervention should be resisted.

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

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

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

Although collaboration among competing banks with respect to the interchange fee should be permitted under the antitrust laws, any expansion of the range of cooperative action should be viewed with healthy skepticism. Thus antitrust and banking authorities should be alert to ensure that the number of payment systems is as large as the attainment of scale economies permits. Though unbridled autonomy within a system cannot be attained, unbridled rivalry between a multiplicity of systems should be encouraged.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

11 *See* Sec. III *infra*.

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

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

14 The use of checks for local payments accelerated after the Revolution. There is substantial evidence of the use of checks in the nation's commercial centers before the creation of the first United States Bank in 1791. Id. at 43. Spahr estimated the amount of check use in America by examining the relation between deposits and currency in circulation. Deposits passed bank note currency in 1855. *Id*. at 60. In 1867 the public held $1.20 in deposits for every dollar of currency and, by 1872, held $2.00 for every dollar of currency. After 1880 the ratio began a long-term climb; it was twelve to one in 1929. Milton Friedman & Anna Schwartz, A Monetary History of the United States 16 (1963).

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

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

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

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

19 Although there is a consensus that the draft was the principal means by which a buyer in the country paid a long-distance debt during the early part of the nineteenth century, there is disagreement about the duration of the practice. Thatcher C. Jones, Clearing and Collections 172-74 (1931); *Testimony on Par Collection of Checks: Hearings on H.R. 12379 Before the House Comm. on Banking and Currency*, 66th Cong., 2d Sess. (1920), indicates the importance of the use of drafts up until the 1890s. But Claudius B. Patten, writing on the mid-1880s, stated that although the use of drafts was common thirty to forty years previously, "Nowadays no country trader, no matter whether he is located in Deadwood or St. Augustine, thinks he is in fashion unless he 'pays' his New York or Boston bills by sending there his individual checks on his local bank, which gets all the advantage of his deposit until the checks come around for collection from the city banks, which have given their dealers immediate credit for them, and made no charge for their collection." Claudius B. Patten, The Methods and Machinery of Practical Banking 1100-01 (11th ed. 1902).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

42 Section 19 of the Federal Reserve Act specified the reserve requirements of member banks. The requirements were substantially lowered by the Act of June 21, 1917, ch. 32, Pub. L. No. 25, § 10, 40 Stat. 239. Member banks in central reserve cities were required to maintain reserves of 18 percent against demand deposits (decreased to 13 percent) and 5 percent against time deposits (decreased to 3 percent). Member banks in reserve cities were required to carry reserves of 15 percent against demand deposits (decreased to 13 percent). The reserves of country banks were fixed at 12 percent for demand deposits (decreased to 7 percent) and 5 percent for time deposits (decreased to 3 percent). The reserve requirements were lowered to stimulate membership in the Federal Reserve System. *See* Federal Reserve Bank of Richmond, Letter No. 5, Apr., 1922, reprinted in Wright, *supra* note 15, at 391-404.

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

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

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

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

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

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

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

50 *Id*.

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

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

53 Federal Reserve Bank of Richmond, Letter No. 6, May 1922, reprinted in Wright, *supra* note 15, at 410-12. This tactic of going to the window of the noncomplying bank and demanding full payment had been used before as a means of achieving a system of par clearance. The Suffolk Bank System in the 1820s (see Justice Story's decision in *Suffolk Bank v. Lincoln Bank*, 22 Mass. 106 (1827)) and the Country Checks Department of the Boston Clearing House in the 1890s (**see** note 32 *supra*) both used the same tactic to force par clearance. The Suffolk Bank System was primarily designed to prevent the discounting of bank notes. *See* Spahr, *supra* note 13, at 73-78, 126-29; Federal Reserve Bank of Richmond, *supra* note 15, at 379.

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

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

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

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

58 *Id*. at 256-90.

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

60 *Farmers & Merchants Bank v. Federal Reserve Bank of Richmond*, 262 U.S. 649 (1923).

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

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

63 *Id*.

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

65 *See Broadcast Music, Inc. v. Columbia Broadcasting Sys., Inc.*, 441 U.S. 1 (1979); *Continental T.V., Inc. v. GTE Sylvania Inc.*, 433 U.S. 36 (1976). However, the Supreme Court has on occasion failed to recognize the significance of maximum price fixing where the product has joint-demand characteristics. *See Albrecht v. Herald Co.*, 390 U.S. 145 (1968). *See also* Frank H. Easterbrook, *Maximum Price Fixing*, 48 U. Chi. L. Rev. 886 (1981).

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