# The Organization of Private Payment Networks

John A. Weinberg

One of the key roles banks have traditionally played is in the execution of payments among participants in the economy. Liabilities issued by banks, such as demand deposits, are a primary means of payment. The widespread acceptability of such private liabilities requires a reliable method for the settlement of such obligations. In a world where people and economic activity are dispersed in space and time, settlement often requires a method for communication between locations where purchases of goods take place and those where payment liabilities are issued. That is, the use of bank liabilities as means of payment requires the support of an interbank network for clearing and settlement.

Throughout U.S. banking history, such multibank networks have played important roles. In New England during the Free Banking period (1836–1863), the system that was centered around the Suffolk Bank in Boston widened the area over which many banks' notes could circulate at par.[1] In the latter part of the nineteenth century, banks participated in clearinghouses for the clearing and settlement of local checks and had correspondent relationships to handle checks over greater distances.[2] While the Federal Reserve (the Fed) ultimately took over a large part of the clearing and settlement of checks, the banking industry has developed other private, multilateral networks for handling interbank payments. Most notable, perhaps, are the nationwide credit card associations.

Recently, interbank networks have received considerable attention. The ongoing growth and consolidation of Automated Teller Machine (ATM) networks has stimulated discussions in the academic and public policy communities of possible antitrust issues raised by such large joint ventures of banking organizations. Much of the discussion of possible public policy concerns regarding

This article has benefited greatly from the comments of Tom Humphrey, Jeff Lacker, Pierre-Daniel Sarte, and John Walter. The views expressed herein are the author's and do not represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] See, for instance, Calomiris and Kahn (1995).

[2] A recent description of check clearing in the nineteenth century is found in Gilbert and Summers (1996).

coming forms of electronic money centers on the network characteristics of these instruments.

This article takes the position that networks are fundamental to the role played by intermediary institutions in the payment system. Clearing and settlement, as the means of managing the financial relationship among individuals or institutions across time and distance, are inherently network services. The characteristics of network services have important consequences for the industrial organization of the payment system.

Arrangements for clearing and settlement of payments, whether private or public, involve an effect that is sometimes referred to as a network externality. Broadly, the private value to an individual of belonging to a network increases with the number of endpoints to which that network connects. Put differently, the private value an individual derives from participating in a network is that the individual can communicate with the other network members. At the same time, the individual's participation creates value for other members by adding to the number of endpoints.

There is an important difference between network externalities and other forms of externality. Perhaps the most common textbook externality is pollution; the economic activity of some individuals may produce pollutants that affect a much broader set of people. While individuals make choices about participation in the activity that generates pollution, they may have little choice as to whether to be affected by pollution. In the extreme case of the greenhouse effect, for instance, it may be impossible to avoid incurring the costs of pollution. Pollution is an external cost that has effects beyond the set of people engaged in the polluting activity. Network externalities, on the other hand, are more self-contained. An individual's decision to subscribe to a network creates external benefits for other subscribers by increasing the size of the network. Notice, however, that in order to enjoy the effects of the externality, one must join the network. The benefits to network participation, although partially external to the individual participant, are entirely internal to the network as a group. Since the group is composed of individuals engaged in mutually voluntary exchange with one another, one would expect the group's organization and pricing arrangements to take account of the "external" benefits associated with an individual's participation.

In a payment network, the value of communication between two endpoints is determined by the pattern of commerce. People at location A will place a high value on being in a payment network with people from location B if there is a high volume of commerce between the two locations. Since locations can vary widely in the sets of places their people go to shop, there can be variety among endpoints in both the private value of network participation and the external value that an endpoint creates for others through its participation. For a network to be sustainable, then, its services must be priced so that all of its intended members have an incentive to join. For a network to be efficient, it

must include all endpoints for which the total value of participation (private plus external value) exceeds the resource cost of participation, and only those endpoints.

This article proposes that the two standards of sustainability (which is defined more precisely below) and efficiency can form the basis of a positive theory of private, multilateral clearing and settlement arrangements.[3] Such a theory is quite distinct from the conventional view that network effects, as a form of externality, are a common source of market failure. This conventional view arises from the analysis of the behavior of network participants under the assumption that key organizational features of networks are exogenously fixed. By contrast, the theory presented below treats organizational arrangements as the endogenous outcomes of interactions among participants. Understanding the differences between these two theoretical perspectives is essential for understanding the role of central banks or other public entities in such activities.

The next section presents an abstract model of a network, gives some possible payment system interpretations, and shows how the essential network characteristics provide implications for network organizations. The following sections discuss some of the elements of a more general (and complete) model and apply some of the insights of the analysis to some historical and contemporary payment network issues. In particular, Section 3 discusses check clearing prior to the founding of the Federal Reserve and current issues involving ATM networks. In the former case, many observers have argued that check clearing was inefficient, as evidenced by the fact that checks sometimes followed very indirect routes in proceeding from initial deposit to final clearing. In the latter case, the use of surcharges (charged by an ATM-owning institution to depositors from other institutions) has been cited as an inefficient exploitation of market power in private networks. Section 3 will present the argument that the empirical facts of both these cases are consistent with a theory that predicts efficient private network organizations.

## 1. AN ECONOMIC MODEL OF A NETWORK

While network effects have often been said to be present in a variety of industries that are not explicitly networks, the focus here is on explicit networks.[4]

[3] Sharkey (1985) and Henriet and Moulin (1996), for instance, follow this approach to the theory of network structure and pricing.

[4] For instance, some authors have suggested the presence of such an externality in the market for personal computer operating systems; from a given set of alternatives, buyers prefer to have the system that is more widely chosen. In this case, the externality comes from the indirect effect that a system's popularity has on the likely availability of application software. Indeed, one might argue that a network effect is present in the retail grocery industry; the value to consumers of shopping at a larger store (or chain) might be enhanced by the store's ability to attract a wider set of suppliers, thereby offering the shopper greater variety. For a general survey of network effects, see Economides (1996).

In particular, the focus is on two-way networks where the underlying service for which people have a demand involves transmissions between two particular individuals. One can think of such a transmission as communication and the underlying product as a message sent from one specific individual to another. The model created below specifies the communication and consumption opportunities available to people in the economy.

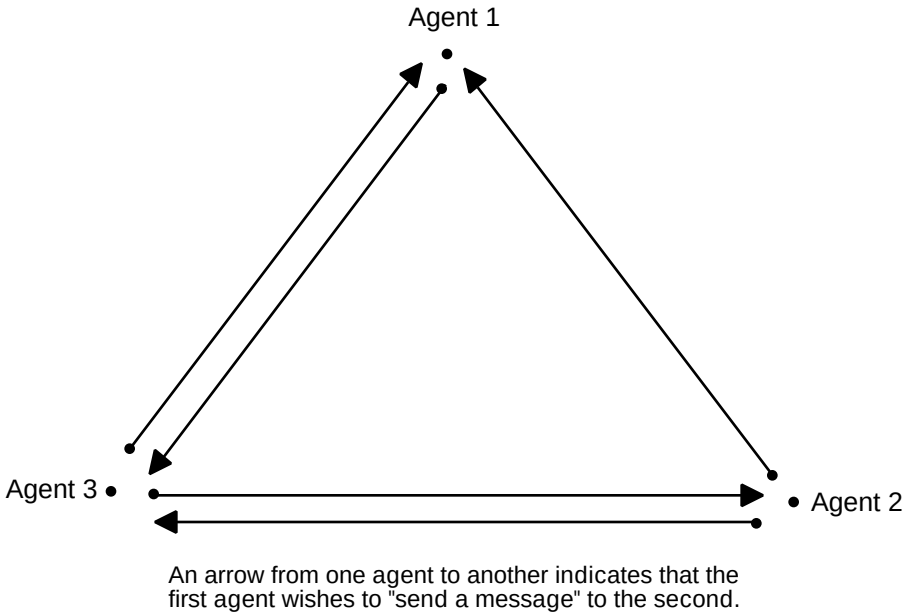
**An Economy with Consumption and Communication**

To create an economic model of a network, imagine a set of individuals, each of whom lives at a distinct location, separated from the others. One can imagine any number of individuals, and in general one might denote the set of individual agents as $\{1, 2, \ldots, N\}$. The key insights from the analysis can be gained from a simple example with three agents. Each individual derives utility from sending messages to some subset of the other individuals. Again, there is a wide array of possible patterns of desired communication. In particular, there may be heterogeneity in the overall value agents place on communication. Some agents may desire to send messages to a large number of recipients, while others may wish to communicate with only a few. The pattern of desired messages that is represented in Figure 1 has this feature. Agent 1 desires only to send a message to agent 3. Agent 2 derives utility from sending messages to agents 1 and 3, while agent 3 would like to send to agents 1 and 2. It is useful to assume that, in addition to communication, agents receive utility from consumption of a generic good and that each agent begins with an endowment of this good.

An agent's preferences for consumption and communication can be stated more formally as follows. Let $J_i$ be the set of other agents with whom agent $i$ wishes to communicate. Hence, in Figure 1, $J_1 = \{3\}$, $J_2 = \{1, 3\}$, and $J_3 = \{1, 2\}$. For any set $J$, use the notation $n(J)$ to denote the number of elements in that set; for instance, $n(J_3) = 2$. The agent receives utility $v$ for each unit of communication, and utility is linear in consumption. If $x_i$ denotes agent $i$'s consumption of the generic good and $U_i$ denotes the agent's utility, then $U_i = n(J_i)v + x_i$.

A model must also specify the technology available for communication. In particular, suppose there are two ways to communicate. A message can be sent by direct, bilateral communication at a cost (to the sender) of $c_0$, in units of the generic good. Alternatively, agents can buy access to a network. A network is a set of "connected" agents. If an agent is connected to a network, then messages to any other agent connected to the same network are costless. The cost, again in units of the generic good, of connecting an individual to a network is $c_s > c_0$. A more general specification of network costs might include a fixed (infrastructure) cost, a cost that is variable in the number of agents connected (like $c_s$ per connected agent in the present specification) and a cost per communication (assumed zero here). The essential feature that is

**Figure 1 A Simple Example of Demand for Network Services**




captured by the simple specification given here is that, by expending resources to connect to one another, agents can reduce their costs per unit of communication. If we assume that the (utility) value of sending a message ($v$) is at least $c_0$, and that the agents' endowment of the consumption good is at least $2c_0$, then each agent will send all desired messages even in the absence of a network connection (the most any agent would spend on communication, if all communication was bilateral, is $2c_0$).

An efficient network is one which includes all agents whose private and external benefits of membership exceed the cost of connection and includes no other agents. To state this definition formally, one needs some additional notation. Let $S$ be a possible network. That is, $S$ is some subset of the agents. This set might also be termed a coalition. The private value to individual $i$ of being a member of the connected set $S$ depends on the number of other agents in $S$ with whom $i$ wishes to communicate. Letting $J_i^S$ denote this set, the relevant number of connected agents for $i$ is given by $n(J_i^S)$. Formally, $J_i^S$ is the intersection of the sets $S$ and $J_i$. Network membership does not change the actual amount of communication in which $i$ engages; it is still worthwhile to communicate bilaterally with all who are in $J_i$ but not in $S$. Therefore, the gross private benefit of membership (gross of connection costs) is the savings

in bilateral communication costs. For agent $i$, this value is $n(J_i^S)c_0$, since agent $i$ wishes to send a single message to each of the other agents with whom he communicates.

In addition to the private value, agent $i$'s membership in $S$ creates value for other members of $S$. Define $H_i$ as the set of agents who wish to communicate with $i$. Hence, in Figure 1, $H_1 = \{2,3\}$, $H_2 = \{3\}$, and $H_3 = \{1,2\}$. Agent $i$'s membership in $S$ creates value for all agents in $H_i$ who are also in $S$. These agents are denoted $H_i^S$, the intersection of the sets $H_i$ and $S$. The external value created by $i$'s membership in $S$ is $n(H_i^S)c_0$. With the private and external values of membership specified, one can state the following definition.

Definition 1: An efficient network is a set of agents $S^*$ such that[5]

(1) $[n(J_i^S) + n(H_i^S)]c_0 \geq c_s$ for all $i$ in $S^*$, and

(2) $[n(J_i^S) + n(H_i^S)]c_0 < c_s$ for all $i$ not in $S^*$.

Condition (1) states that if agent $i$ is in the network $S^*$, then the private-plus-external benefits from $i$'s membership equal or exceed the connection cost. Condition (2) states that if private-plus-external benefits are less than the connection cost, then agent $i$ is not in the network. Under the additional assumption that $3c_s < 5c_0$ (costs of universal connection are less than total communication costs in the absence of a network), an efficient network for the example of Figure 1 is one that includes all three agents.

### Behavior of Agents in the Economy with an Exogenous Organizational Structure

Given an economic environment such as the one just described, how does one predict an outcome? In particular, what sort of network will emerge? Will it be efficient? How will agents share the costs and benefits of network connection? One approach to these questions is to assume a certain form of competition among potential networks. That is, one might assume a particular game played by the participants in the economy. To this end, it is useful to assume that there is an additional set of agents who have access to the network technology and derive utility only from the consumption good. These agents compete by offering network services to the agents who have a demand for communication. For simplicity, suppose that the services of these potential network providers are incompatible; each provider's network is unique and cannot be connected to other providers' networks. The "rules of the game" dictate the types of offers

[5] To be precise, the notation in Definition 1 should specify the particular network $S^*$. For instance, $J_i^S$ should be $J_i^{S^*}$. When there is only a single network, as in Definition 1, suppressing the extra notation introduces no ambiguity. Further, for this economic environment the assumption of a single network is without loss of generality.

the sellers are allowed to make as well as the allowed responses by network users.

One possible game through which network providers could compete proceeds as follows. First, each potential provider announces an access price and stands ready to provide access to all comers at that price. Next, agents choose whether and from whom to buy access. Finally, communication and consumption take place. This game will be referred to as the *uniform price-setting game*. This name reflects the fact that sellers are not permitted to price discriminate by offering different prices to different buyers. The predicted outcome of a game is its Nash equilibrium, which is a set of strategies (price offers by providers and network choices by buyers) such that each player's (agent's) strategy is (privately) optimal, given the strategies of all the other players. Sellers in this game essentially bid for the right to provide network services to all who sign up. Since each potential provider is just as capable as all the others, price competition will tend to drive profits to zero. With uniform price setting, this competition can lead to extreme results, as in the following.

> Result 1: In the equilibrium of the uniform price-setting game for the environment described by Figure 1, no agent purchases network access, so all communication is bilateral.

To see that this result is true, note first that the connection fee must be at least $c_s$, since sellers have no incentive to sell at a loss. Agent 1, however, will not pay $c_s$ to join the network, since joining saves him at most bilateral costs of $c_0$ (if agent 3 is in the network). Without agent 1 in the network, connection is not worth $c_s$ to the other agents, so the equilibrium network is empty.

This result is a stark example of the general finding that under certain forms of competition equilibrium involves networks of inefficient size.[6] For the uniform price-setting game, such a result arises whenever there is at least one agent for whom the greatest possible private benefits of connection are less than the connection cost but for whom the private-plus-external benefits exceed costs. Such an agent is one who has a relatively low personal demand for communication but with whom many others wish to communicate (an agent for whom $n(J_i)$ is small and $n(H_i)$ is large).

### An Alternative Model of Behavior: Sustainability

Results like Result 1 are useful for understanding the nature of competition in network markets. Their usefulness is limited, however, because they typically apply to specific games. The variety of games that might be played is virtually endless, even for the very simple economy of Figure 1. Would other games yield different results? In particular, are there games for which the equilibrium

[6] See Economides (1996) for a survey of such results.

network is efficient? The process of searching over all possible games is an impractical approach to the economic analysis of such environments. An alternative is to look directly at possible outcomes (allocations of network services and the consumption good) and ask which outcomes are sustainable, in a well-defined sense. In particular, suppose a certain allocation has been tentatively agreed upon by the agents in the economy. That allocation is sustainable if no subset of the agents can make themselves better off by allocating their own endowed resources among themselves. In the present environment, such a deviation from a proposed allocation would involve the subset of agents, or coalition, forming its own network and communicating with all other agents bilaterally; this coalition might also choose to reallocate its members' endowments of consumption goods, net of connection costs.[7]

To state this sustainability property more formally, an allocation in this economy can be defined as a network $S$ and a payment for each agent, $p_i$, toward covering connection costs.[8] In considering possible deviations from proposed allocations by coalitions, it is sufficient to consider a coalition of any size that deviates to form a single network; if there were an incentive for a coalition to form two distinct networks, then each network would have its own incentive to form, regardless of whether the other forms. Letting $p = (p_1, p_2, \ldots, p_N)$, we can now define sustainability as follows.

Definition 2: An allocation $(S, p)$ is sustainable if there does not exist a network $S'$ and a payment allocation $p'$ such that

(1) $p'_i + n(J_i - S')c_0 \leq p_i + n(J_i - J_i^S)c_0$ for each $i$ in $S'$, with strict inequality for some $i$ in $S'$, and

(2) $\sum_{i \text{ in } S'} p'_i \geq n(S')c_s$.

In the above definition, the notation $A - B$ for two sets $A$ and $B$ means all elements of $A$ that are not in $B$ (the complement of $B$ on $A$). The first condition states that each agent in the deviating network $S'$ is made better off by joining that network (and at least one is made strictly better off). An agent is made better off if his total outlays for bilateral communication and connection payments are reduced. The second condition is that the network collect enough in payments to pay for connecting all of its members.

One important fact to note about sustainability is that a sustainable allocation must be efficient. If it were not, a coalition consisting of an efficient network could form and arrange an allocation satisfying the two conditions above. In the case represented by Figure 1, a sustainable allocation will involve

[7] A sustainable allocation is defined here as an allocation in the *core* of the economy.

[8] In principle, one can allow for the existence of more than one network, although in this environment any allocation with more than one network is either equivalent to or (Pareto) dominated by an allocation with a single network.

a single network consisting of all three agents. The remaining task is to determine how the costs of connecting the agents ($3c_s$) should be shared among the three. As we have already seen, charging each agent $c_s$ does not work; agent 1 can be charged no more than $c_0$. The remaining cost ($3c_s - c_0$) must be covered by charges to agents 2 and 3, with the restriction that neither can be charged more than $2c_0$, the cost to each of sending all messages bilaterally. Hence, one possible cost allocation is for agent 1 to pay $c_0$, agent 2 to pay $2c_0$, and agent 3 to pay $3(c_s - c_0)$. These prices give no agent an incentive to leave the network and send all communications bilaterally. The prices also leave no incentive for any pair of agents to form a separate, two-agent network and communicate with the third bilaterally. Hence, with these prices the efficient network is sustainable. This result is summarized below.

Result 2: In the case of Figure 1, a sustainable allocation has an efficient network $S = \{1, 2, 3\}$ and any payments $(p_1, p_2, p_3)$ satisfying $p_1 \leq c_0$, $p_2 \leq 2c_0$, $p_3 \leq 2c_0$, and $p_1 + p_2 + p_3 = 3c_s$.

**A Sustainable Allocation as the Outcome of a Game**

Returning now to the question of competitive games played by potential providers of network services, does there exist a game under which the equilibrium network is the efficient network? Suppose, as before, that a large number of (incompatible) network service providers compete for subscribers. Instead of requiring that competition be only in the form of nondiscriminating access prices, suppose that the sellers can make any type of price offer they wish. That is, price offers can be in the form of a distinct price for each buyer. Refer to this game as the *perfectly discriminating price-setting game*. Equilibrium prices for this form of competitive bidding correspond to the sustainable cost allocations specified above.[9]

Result 3: Equilibria of the perfectly discriminating price-setting game are sustainable allocations.

To see this, suppose a seller has offered a set of prices satisfying the conditions stated in Result 2. Can any other seller offer prices that win customers from the first and yield a profit? In order to attract any individual buyer, a competing offer must give that buyer a lower price. In order to cover costs, however, the payment from at least one other buyer must be raised, so it is impossible to attract more than one buyer. If all buyers do not join, network connection is not attractive. Hence, a payment allocation satisfying the sustainability conditions cannot be undercut. On the other hand, any set of prices that

[9] There are many equilibria corresponding to many core allocations. In all of them, agent 1's access price is no greater than $c_0$, agents 2 and 3 face prices no greater than $2c_0$, and all agents connect to the network.

leaves a seller with strictly positive profits can be undercut. Accordingly, the sustainable allocations correspond with the set of equilibria.

The most notable feature of sustainable pricing arrangements is that, in order to support an efficient network, they require the subsidization of agent 1's connection by the other two agents; agent 1's connection fee must be less than the resource cost of connecting him. The other agents are willing to cover the remainder of the cost, because agent 1's (social) value to the network exceeds the (private) value he places on network access. This example illustrates a general point about arrangements that support efficient networks in environments in which agents are heterogeneous in the way they value network participation. Benefits (and possibly the costs) of network participation have a collective component. The key to sustaining an efficient network is in the distribution of these collective benefits and costs. This distribution must respect the capability of agents to leave the network, either individually or in groups. In a setting with heterogeneous agents, it can quite easily arise that the appropriate distribution of costs and benefits require that different agents pay different prices for essentially the same service.

The pricing arrangements that satisfy the conditions in Result 2 involve perfect price discrimination; they require that prices be tailored for each individual buyer of network services. Perfect discrimination is not always feasible. If, for instance, there is uncertainty about demands for network services and an individual's true demand characteristics are private information, then prices cannot be as finely targeted as in the above example. In this case, private information imposes further constraints on attainable allocations. It is possible to incorporate such constraints into the notion of sustainable arrangements. In such settings, pricing arrangements are likely to involve a less perfect form of price discrimination. For instance, prices for network services may be tied to observable characteristics or actions that are correlated with true demand. In an environment similar to the one discussed in this section, access prices that vary with the amount of communication might be able to achieve the desired price discrimination in a way that allows privately informed buyers to self-select among alternative pricing options.

## 2. ELEMENTS OF A GENERAL PAYMENT NETWORK MODEL

The model and example of the previous section were specified in terms of a generic communication service. The same sort of network structure, however, can arise in a model that is specified in such a way as to capture important aspects of payment system markets. Any noncash payment mechanism is a communication network in a fundamental sense. An instrument presented in payment for goods or services is an instruction to transfer monetary value from the buyer's to the seller's ownership. Execution of such an instruction requires

communication between the point of sale and the location or institution at which the buyer's value is held. This section sketches some of the ingredients of a general payment network model. The key point is that the private and external values to individuals of being connected to a payment network depend on the underlying pattern of commerce.

**An Economy with Payment Services**

As in the above section, suppose that there is a set of $N$ distinct locations at which agents live and economic activity takes place. Unlike the previous section, suppose that there is a large number of agents living at each location. Agents consume two types of goods: a generic good and location-specific goods. Different people have different preferences for location-specific goods. In particular, each agent desires the specific good from exactly one location. One might, for instance, denote by $\phi_{ij}$ the fraction of agents from location $i$ who wish to consume the specialized good at location $j$. These fractions determine the economy's pattern of commerce. Agents travel from their home location to the locations at which they wish to consume and purchase location-specific goods with claims on amounts of generic goods (or with the generic good itself). In some environments, one might also imagine that these transactions are made using government-issued fiat currency.

Making transactions across locations is costly. This cost might arise from a number of frictions. If debt claims are created in the purchase of location-specific goods, then there may be costs associated with communicating information about these claims across locations or in making final payment. If buyers carry the generic good with them to make purchases, there may be transportation costs or other losses incurred on the way. Finally, if traveling for consumption takes time and buyers carry non-interest-bearing currency for transaction purposes, then there may be a seigniorage cost associated with location-specific consumption. The specific nature of the costs depends on the details of the economic environment.

As above, suppose that there is a technology for connecting locations in a network for the purpose of clearing and settling payment obligations. While connection may involve a fixed cost, the variable cost of making transactions among connected locations is lower than between locations that are not connected. For instance, if payment for location-specific goods requires shipment of generic goods, a network that allows multilateral communication and calculation of net obligations may economize on shipping costs. Alternatively, in a monetary economy, a network that allows people to substitute debt claims for currency may allow agents to save on seigniorage costs.

The value to agents at a particular location of being connected to the network depends on which other locations are connected. In particular, agents at location $i$ place a high value on being in a network that includes locations $j$ for which the fractions $\phi_{ij}$ are large. By the same token, a location at which

many people want to shop ($j$ such that $\phi_{ij}$ is large for many $i$) is one that brings a high external value to any network it joins. Hence the notion of sustainable pricing of network services, as presented in the previous section, will imply that these popular locations receive preferential treatment in the pricing of network services. As is the case for location 1 in the example of Figure 1, there could easily be locations for which the private value of network services is small while the external value is large. These would be locations that attract many consumers from elsewhere but whose residents consume mostly their own location-specific goods. Pricing to support an efficient network could require that such locations pay less than the cost of connecting them, as is the case in the example.

Payments-related models that have the features outlined in this section will have the same implications as the example from the previous section. Network industries will tend to be organized in ways that achieve sustainable allocations. This means, first, that there is a tendency toward efficient network structures. Second, the sharing of the costs and benefits of network organizations must respect the ability of participants to form or join alternative organizations. Accordingly, network members who bring large external benefits to other members may need to receive a share of the net benefits that appears to be out of line with those members' own use of the network services. Such an impression mistakenly focuses only on an individual's private benefits of network participation and not the benefits that an individual brings to other participants.

**Barriers to Competition**

An important maintained assumption in the foregoing discussion is that any agent or group of agents that is dissatisfied with an arrangement is free to pursue an alternative. That is, there are no barriers to entry. Various types of barriers might arise in economic environments. For instance, if all agents do not have access to the same technological capabilities, then it might be difficult for agents that are dissatisfied with their network services to set up or seek out an alternative network. Also, there may be investments in network provision or participation that, once made, represent sunk costs. A sunk cost is a cost that cannot be fully recovered. In this case, an incumbent network would have a cost advantage over a competitor; while the competitor must incur the sunk costs, those costs are no longer part of the incumbent's decision calculus.

Other barriers to competition might arise from legal restrictions. For instance, if sellers were to face a legal prohibition of price discrimination, then the types of network services arrangements they could offer would be sharply limited; as seen above, price discrimination can be essential for the efficient provision of network services with heterogeneous buyers. Other legal barriers might take the form of restrictions on which particular sellers can offer which particular services.

A final form of potential barrier that is worth noting arises from the behavior of sellers themselves. An incumbent seller of network services might attempt to impose rules on its buyers that make it difficult for them to switch to a competing service. The possibility of restrictive rules set by network providers has been a subject of interest in recent policy discussions concerning ATM network mergers.

If there were barriers to competition, then an inefficient network structure could persist. In such a case, can public policy intervention improve on private market performance? The answer depends on the source of the barriers. In the United States and other developed economies, enforcement of antitrust laws is in part intended to guard against the anticompetitive use of restrictive rules by sellers in their contracts with buyers. In cases where a barrier to competition is the result of a legal restriction on the behavior of market participants, such restrictions might have other public purposes. Here, as in the case where barriers may have technological sources, it may be difficult or impossible for public intervention to remove the barriers. In these cases, an incumbent provider might extract monopoly rents, for instance, by inefficiently limiting network size.[10] In such cases regulation of the pricing and product offerings of an incumbent seller might be useful in promoting network efficiency. Governments in many economies have traditionally taken this approach to telecommunications markets.

## 3. TWO APPLICATIONS

One can apply the logic presented above to a number of actual payment network examples. This section will provide a brief discussion of two such examples, one historical and one current. The historical example involves the process of clearing checks in the United States in the period prior to the founding of the Federal Reserve System. The current example involves the growing geographical reach of multibank ATM networks.

### Check Clearing before the Federal Reserve

By the late nineteenth century, checks had already become a dominant form of payment. As the banking industry was highly fragmented, a significant portion of all checks were interbank checks. Clearing of checks between the depositor's and the check writer's bank occurred in one of a number of ways.[11] A bank

[10] Of course, even a monopoly immune to competition will not necessarily produce inefficient results. If the monopolist has sufficient ability to price discriminate, monopoly behavior can approach full efficiency. In this case, the monopolist's rents come at the expense of consumer welfare but not at the expense of total welfare.

[11] For a recent discussion of pre-Fed check clearing see Gilbert and Summers (1996). A classic detailed account is found in Spahr (1926).

holding checks drawn on accounts at another bank could present the items directly, in person, at the paying bank. By law, the paying bank was required to make payment on such checks without imposing any presentment fee. On the other hand, checks that were presented through the mail could be subject to a fee imposed by the paying bank. Banks could also clear checks through the services of an intermediary, or correspondent bank. Banks with a correspondent relationship might have entered into a mutual agreement to accept checks from one another without imposing fees.

At a time when bank branching was limited by law, correspondent relationships were particularly important for clearing checks in cases where the paying bank was relatively distant from the bank in which the check was initially deposited. If both banks had correspondent relationships with the same intermediary bank, then collection of the item could proceed free of fees. In this sense, any two banks that were connected by a chain of correspondent relationships belonged to a network. How was the composition of such networks likely to have been determined? The value to a bank of belonging to a network depended on the frequency with which the bank received checks drawn on other members of the network. Further, if there were particular institutions that had relatively frequent and large volume interactions with many other banks, then such institutions would naturally serve as central intermediaries in a correspondent network. For instance, a city bank might deal with a number of country banks in the surrounding region. The city bank might, in turn, maintain relationships with banks in other cities that serve as correspondents for their regions. This organization of a check-clearing network could economize on the costs of shipping checks. A small bank in a remote town could, for instance, send a single shipment of all its out-of-town checks to its correspondent, with the links between larger correspondents serving as "trunk lines."

In a correspondent network like that described above, consider the problem faced by a bank that receives a check drawn on an institution with which it rarely deals. The receiving bank could send the item directly to the paying bank. In this case, however, the paying bank might charge a fee for presentment. Alternatively, the receiving bank could send the item, along with its usual shipment, to its correspondent bank. Then, through a chain of correspondent relationships, the check might ultimately be paid at its par value (with no presentment fee). This indirect alternative has two potential sources of savings. First, presentment fees might be avoided. Second, there may be savings on the costs of transporting checks. The marginal cost of adding an item to a routine shipment is virtually zero, certainly smaller than the postage cost of sending a single item directly. Indeed, similar economies may have been available at the receiving end of check shipments. A paying bank may have found it more convenient and cost effective to receive and process bundles of checks sent by intermediaries with whom it had a standing relationship.

### *The Circuitous Routing of Checks*

The history of check clearing during the period that preceded the founding of the Fed contains examples of checks traveling over circuitous routes to get from the banks in which they were initially deposited to the paying banks. A bank of first deposit, for example, might have sent a check to its correspondent located to the east even though the paying bank was located to the west. There are two possible interpretations of such examples. On the one hand, such instances might provide evidence of an inefficiency created by paying banks' ability to assess presentment fees. On the other hand, such cases could be consistent with the operation of an efficient correspondent network. The pattern of links in the network (correspondent relationships) was determined by the usual pattern of commerce. The occasional circuitous route for check clearing resulted simply because there were occasional exceptions to the usual pattern of commerce. Given the existing links, it was efficient to send these occasional items together with routine shipments. Indeed, in this view, it is possible that presentment fees reinforced network efficiency by reducing the incentive for individual banks to bypass the network.

The second interpretation is consistent with the analytical framework suggested above. Under this interpretation, presentment fees may have actually reinforced efficient check-clearing relationships by discouraging the direct presentment of occasional, solitary items. Those few items that were sent directly and on which presentment fees were paid are likely to have been items for which the bank of first deposit could not foresee a sufficiently reliable chain of correspondent relationships. That is, these were items for which the bank that was due payment had little alternative to direct presentment (through the mail). Accordingly, such items would constitute a market segment (in the market for clearing by direct presentment through the mail) with relatively inelastic demand. The efficiency cost of charging a high price (above marginal cost) to market segments with inelastic demand is relatively small; with inelastic demand, quantity purchased does not decline much as the price is raised. In other words, presentment fees allowed paying banks to price discriminate between institutions with good alternatives to direct presentment and those without such alternatives. It is entirely possible that the effects of such discrimination were primarily distributional. While some buyers gain at the expense of others, price discrimination typically increases the exploitation of gains from trade relative to uniform pricing in the presence of market power.[12]

### ATM Networks

An ATM transaction, like a check transaction, can require interbank clearing and settlement. When the holder of an ATM card issued by one bank makes a

[12] Each paying bank had some market power in the sense that it was the only bank that could provide final settlement of a check.

withdrawal at an ATM owned by another bank (or perhaps by a nonbank business), the transaction must be cleared by communicating information about the withdrawal to the card-issuing bank and settled with an appropriate transfer of funds from the issuer to the ATM owner. In recent years, multibank ATM networks have become increasingly important in allowing cardholders to access their funds at ever widening sets of locations.[13] A regional ATM network is usually organized as a joint venture owned by some or all of its participating banks. The network typically has a brand name that is placed on its members' machines and cards.

Membership in an ATM network is valuable to a bank mainly because access to the network's set of ATM locations enhances the value of the ATM services the bank offers its depositors. Clearly, membership is more valuable the more extensive the network's set of locations. A bank's membership also brings "external" benefits to other members by adding to the set of locations. Some locations, however, are more valuable than others. For instance, banks (and their customers) may place a high value on having access to an ATM at a vacation resort. Such a location may be one for which the external benefits of network participation are greater than the private benefits to the owner of the particular ATM. The theory presented above suggests that a sustainable network will need to price its services or share its profits so as to allow the ATM owner to realize some of the external benefits of its participation.

### *Surcharges*

There are a number of ways in which network arrangements could induce participation of institutions that bring large external benefits. For instance, if the network imposes membership fees, lower fees could be set for members with more desirable locations. Similarly, if the network is organized as a joint venture, then arrangements for profit sharing could conceivably reflect differences in the values of members' locations. Since the values of locations are likely to be related to the intensity of ATM use, prices based on transactions might also be used for allocating network costs and benefits. In ATM networks, when a cardholder from one member uses the ATM of another member, the cardholder's bank pays an "interchange" fee to the ATM owner. Since this fee is typically set by the network and is usually a uniform, per-transaction charge, its flexibility for cost and benefit sharing is limited. Owners might also be able to realize some of the network benefits of desirable locations by imposing a transaction fee directly on the cardholder. Cardholders' willingness to pay such fees, known as surcharges, would depend on the value of having access to cash at the particular location and on the degree of competition for cash access services at that location.

[13] A description and history of ATM networks are found in Baker (1996).

Many regional ATM networks, as well as the national networks owned by the Visa and Mastercard associations, once had implicit or explicit agreements among their members banning surcharges. These restrictions were ultimately challenged, often by owners of ATMs at high-value locations.[14] The debate over surcharges centers on two opposing interpretations of the role such fees play in the market. One interpretation holds that a ban on surcharges prevents ATM owners from abusing the monopoly power they gain from having desirable locations. The other interpretation is suggested by the arguments presented in this article; surcharges support the formation of efficient networks by allowing participants who bring large external benefits to the network to capture some of those benefits. Under this second view, a ban on surcharges is an attempt by network owners to impose a certain distribution of network benefits, a distribution that may not be sustainable.

Does the second view imply that monopoly rents earned by ATM owners with particularly advantageous locations do not create the inefficiency usually associated with monopoly power? Not necessarily. If a network includes some truly unique locations and there is no possibility of entry by a competitor at those locations, then the unique locations are essentially natural monopolies. Banning surcharges does not eliminate the rents from those monopoly positions. Rather, a ban is an attempt to spread those rents among all the banks in the network. If a network includes valuable locations offering equal access for customers of all network members but not for nonmembers, then members can extract rents in the fees they charge their customers for account services that include ATM access. In this case, the rents would be extracted from all customers, even those with no demand for access to the highly valued locations. Allowing surcharges, on the other hand, allows monopoly rents to be collected in a more discriminatory fashion. Such price discrimination can have the effect of reducing the inefficiency from monopoly power.

A second important point about monopoly rents is that an ATM owner's ability to extract rents is limited by competitors' ability to enter monopolized market segments. Hence, the primary public policy concern with market power should be "How is it maintained?" rather than "How is it exploited?" Competitors' incentives to enter markets by placing ATMs at particular locations is greatest when owners can earn location-specific rents. Hence, a ban on surcharges creates a situation in which incentives to engage in location-specific competition are muted.

[14] See for instance *Bank Network News,* September 13, 1995.

## 4. CONCLUSION

The theoretical framework presented above suggests that the organization of networks is driven by the desire of market participants to devise sustainable multilateral arrangements. In both of the cases discussed in the previous section, the framework leads one to interpret observed network structures and pricing arrangements as components of an efficient arrangement. Such an arrangement must take into consideration both the private value that participants derive from the network and the external benefits that participants bring to the network. Hence, in a world with heterogeneous demands for network services, price discrimination and other means of "unevenly" distributing the net benefits of network services can be essential for supporting efficient network structures.

Does this article's framework necessarily imply that all actual networks are efficient and that public intervention can never improve upon the economic performance of a private network? As discussed in Section 3 above, this best-of-all-possible-worlds result holds strictly only when there are no barriers that prevent groups of economic agents from pursuing the alternatives of their choice. The role of public policy, therefore, is to understand the sources of barriers to such choice. There may be some cases in which barriers are technological and cannot be overcome. In these cases, some regulation of pricing practices might be called for. This argument, however, is nothing more than the traditional justification of regulation of a natural monopoly. In other cases, barriers may be created through the rules imposed by incumbent providers of network services in an attempt to preserve market share. Public intervention to eliminate such rules could be beneficial. This argument closely mirrors traditional justifications for antitrust scrutiny of the conduct of firms with large market shares. In all cases, the framework begins with a presumption of efficiency. This presumption is expressed by the question, "If all economic decisionmakers are always free to make alternative arrangements, why wouldn't the arrangement on which they actually agree be efficient?" This presumption seems also to be a good place for public policy to begin.

## *REFERENCES*

Baker, Donald I. "Shared ATM Networks: the Antitrust Dimension," *The Antitrust Bulletin*, vol. 41 (Summer 1996), pp. 399–425.

*Bank Network News*. "New Math Renews Old Surcharge Debate," September 13, 1995, p. 1.

Calomiris, Charles W., and Charles M. Kahn. "The Efficiency of Self-Regulated Payment Systems: Learning from the Suffolk System," *Journal of Money, Credit, and Banking*, vol. 28 (November 1996), pp. 766–97.

Economides, Nicholas. "The Economics of Networks," *International Journal of Industrial Organization*, vol. 14 (October 1996), pp. 673–99.

Gilbert, R. Alton, and Bruce J. Summers. "Clearing and Settlement of U.S. Dollar Payments: Back to the Future?" Federal Reserve Bank of St. Louis *Review*, vol. 78 (September/October 1996), pp. 3–27.

Henriet, Dominique, and Herve Moulin. "Traffic-Based Cost Allocation in a Network," *RAND Journal of Economics*, vol. 27 (Summer 1996), pp. 332–45.

Sharkey, William W. "Economic and Game Theoretic Issues Associated with Cost Allocation in a Telecommunications Network," in H. Peyton Young, ed., *Cost Allocation: Methods, Principles and Applications*. New York: North-Holland, 1985.

Spahr, Walter. *The Clearing and Collection of Checks*. New York: The Bankers Publishing Co., 1926.