FEDERAL RESERVE BANK OF PHILADELPHIA

Ten Independence Mall
Philadelphia, Pennsylvania 19106-1574
(215) 574-6428, www.phil.frb.org

WORKING PAPER NO. 97-19

NETWORK DISECONOMIES AND OPTIMAL STRUCTURE

Sherrill Shaffer
Department of Economics and Finance
University of Wyoming

July 1997

WORKING PAPER NO. 97-19

NETWORK DISECONOMIES AND OPTIMAL STRUCTURE

Sherrill Shaffer
Department of Economics and Finance
University of Wyoming
P.O. Box 3985
Laramie, WY 82071-3985

July 1997

JEL codes: L10, G21, D20

Telephone: (307) 766-2173 Fax: (307) 766-5090 e-mail: shaffer@uwyo.edu

This paper was begun while the author was at the Federal Reserve Bank of Philadelphia. This paper represents the views of the author and does not necessarily represent the views of the Federal Reserve Bank of Philadelphia or of the Federal Reserve System. The author is grateful for helpful comments from Ken Kopecky and colleagues at the Federal Reserve Bank of Philadelphia, and for research assistance from Jim DiSalvo and Louise Berna.

ABSTRACT

This paper explores the effect on costs when firms within an industry must interact with each other in the normal course of business. Such interaction will generally cause the socially optimal scale of each firm to deviate from its minimum average cost scale. In addition, the socially optimal industry structure may be more concentrated than conventional firm-level cost studies would suggest and may also differ from the unregulated (free-entry) equilibrium structure. These concepts, while potentially applicable to several industries, are here made more precise for the banking industry, both theoretically and empirically.

NETWORK DISECONOMIES AND OPTIMAL STRUCTURE

Public policy toward industrial structure has traditionally adhered to the neoclassical assumption that an industry can efficiently supply a growing market at constant marginal cost through the entry of new, optimal-sized firms over time. However, to the extent that firms within an industry must interact with each other, as is true of many service industries, the costs of that interaction may be an increasing function of the total number of firms in the industry. Such interfirm network diseconomies may have at least three results. First, unit costs will rise as the market grows, even if individual firms are at their efficient scale. Second, the socially optimal scale of each firm may deviate from its minimum average cost scale, once these network diseconomies are taken into account. Third, the socially optimal industry structure may be more concentrated than conventional firm-level cost studies would suggest, and may also differ from the unregulated (free-entry) equilibrium structure.

This paper explores these concepts theoretically and empirically in the specific context of the banking industry. Evidence emerges that U.S. banking may exhibit network diseconomies, with important implications for both equilibrium structure and optimal public policy. The empirical results also incorporate what appears to be the first test of agglomeration effects in the banking industry, identifying both localization diseconomies and urbanization economies.

1. Background

Previous studies of network effects have focused largely on issues of compatibility, innovation, consumer demand, and competition (see, for example, pathbreaking studies by Farrell and Saloner, 1985; Katz and Shapiro, 1985, 1986; or the survey by Economides, 1996). The aspects of interfirm networking considered here, by contrast, may be essentially transparent to the consumer and primarily affect a firm's costs. Examples of industries exhibiting interfirm networking in fact or in principle include the postal service, telecommunications, airlines, and banking.[1] A primary characteristic of such industries, relevant to this study and in contrast to many previous studies, is that consumers value *coverage* (access to all endpoints) of an exogenous network or market, rather than the size of that network (number of endpoints) per se. For example, when mailing a letter, writing a check, or making a telephone call, a consumer needs a service provider that can deliver the letter, effect payment, or complete the call to any potential recipient. Similarly, a traveler wants to be able to reach any desired destination.

In the situations considered here, we will assume that the demand for global coverage is either absolute or at least sufficiently inelastic that any equilibrium market outcome will yield global coverage.[2] Given such global access or coverage, the consumer may be indifferent to whether a single provider maintains direct links with all endpoints or instead must interact with one or more other firms to reach a particular endpoint, at least insofar as such interactions are transparent to the consumer. For example, if a traveler must change airplanes en route to a given destination, she may be indifferent to using one airline versus two for the same trip as long as the fare and travel times are the same. Similarly, a consumer mailing a letter may not care whether the postal service chooses to subcontract part of the delivery to third parties, as long as the letter arrives promptly and intact. As long as the *industry* provides global access by some combination of means, the size of the network and the structure of the industry do not affect consumer demand. Therefore, only cost considerations will influence firms' choice of scale and scope or (assuming that competitive pricing can be assured) societal preferences regarding industry structure.

If a firm in one of these industries hopes to avoid the need for systematic interaction with its rivals *as an intrinsic step in the provision of the service*, it must establish its own direct links with every endpoint to compete effectively. In this regard, the problem considered here is similar to the interconnected networks problem of Laffont et al. (1996), in which each firm controls a bottleneck (its own customers) to which rivals must have access, and in which the interconnections render a consumer indifferent to the relative sizes of the firms. However, Laffont et al. focus on the pricing equilibria of such cases, whereas this paper--like Radner’s (1992) analysis of hierarchies--focuses on the resulting characteristics of cost. Because it confronts an industry with the need to choose between markets (interfirm linkages) and hierarchies (intrafirm networks) in serving its clientele, the problem considered here also constitutes an application of Williamson’s (1975) analysis and is thereby related not only to Radner’s (1992) study but also to Neave’s (1991) interpretation of financial industry structure in terms of markets and hierarchies. Neave focuses on the costs and capabilities of governance mechanisms in financial firms and, secondarily (as in his Chapter 5), on asset-side transactions; by contrast, although our empirical section can reflect the various contributions to costs of asset, liability, and governance operations, our theoretical banking model focuses on a subset of liability-side transactions.

Although the focus of this paper is on the cost effects of interfirm linkages, there may well be competitive or pricing effects as well. Firms that choose to maintain direct connections to each endpoint will face their rivals in each geographic market, with behavioral effects that have been analyzed in the literature on multimarket contacts. The alternative structure of maintaining interfirm linkages can reduce the number of multimarket contacts but may have additional competitive effects that have yet to be studied.

Different industries have chosen different structures to solve the linkage problem. UPS and Federal Express have established the capability of delivering to any domestic location, while competing long-distance telecommunication companies must connect with local monopoly switching systems to provide their service. We assume compatibility among interacting firms in such cases, unlike previous studies that treat compatibility as a strategic choice variable.

Exogeneity of the relevant network is crucial to our focus. The technology of providing a service may suffer aggregate diseconomies of scale beyond some level; yet, when the total population grows beyond that level, the industry cannot exclude new citizens from its scope, and therefore cannot limit its overall scale to the minimum-cost level. The postal service and telecommunications companies cannot reasonably restrict their coverage to a subset of the population residing in their coverage areas, nor can a commercial bank limit its acceptance of checks to those written by or to a selected few parties. Moreover, these industries face an exogenously growing population over time. Globalization and the erosion of local geographic market boundaries contribute to a further increase in the effective market size.

It is realistic to expect that diseconomies of scale may characterize one or more of these industries. Casual observation suggests that the postal service may operate in a region of decreasing returns, comparing service level and costs over the past few decades. In the 1950s, letters traveled coast to coast in a day or two at a cost of 3 or 4 cents, with twice-daily home delivery, without benefit of high-speed jet aircraft or high-speed automated sorting technology, and without the consumer inconvenience of ZIP codes. Today, the cost of a first-class postage stamp is twice as high in real terms, the average transit time longer and more variable, the frequency of delivery lower, and postal codes increasing in length every few years (two digits when introduced, then five, now nine)--despite technological progress and privatization to realign incentives. The main variable driving this apparent deterioration seems to be the sheer volume of mail, occasioned by both the 60 percent growth in U.S. population over the period and the increased use of mail for commercial advertising.[3]

In banking, interfirm linkages operate at several levels--on the asset side (for example, loan participations), the liability side (for example, payment transactions), or a mixture of the two (as with interbank lending--the so-called fed funds market). In addition, interbank linkages arise with financial services such as correspondent bank relationships to provide electronic payments, check processing, coin and currency, and securities wire transfer services.

On the liability side, the ratio of interbank payment transactions (so-called "transit" transactions) to "on-us" transactions is an increasing function of the number of banks in the market. (This property will be explicitly calculated in the next section under the assumption of isotropic exchange.) To the extent that the receiving bank and the paying bank must duplicate certain steps in the processing of a single transit transaction, the total cost of that transaction will exceed the cost of an otherwise equivalent on-us transaction. Where cost considerations dictate the use of a clearing facility (such as the Federal Reserve or a correspondent bank) rather than direct presentment between the paying and receiving banks, an additional institution is imposed in the chain with its attendant costs.

On the asset side, loan participations and loan sales generate comparable interfirm linkages. If participating or acquiring banks or their agents perform independent credit analysis, costly duplication of effort is involved; if not, the participating or acquiring banks are exposed to moral hazard from the originating or lead bank. In addition, Broecker (1990) and Nakamura (1993) have identified an indirect cost of multiple-bank markets in terms of declining credit quality and increased loan losses resulting from loans granted to applicants previously rejected at other banks. Structural consolidation is occurring at a rapid pace among U.S. banks (including the very largest) amid claims of analysts and practitioners of potential cost savings, despite the findings of most empirical cost studies that scale economies are exhausted beyond some smaller scale (Berger and Humphrey, 1992b).[4] The analysis below suggests that network diseconomies may potentially account for at least part of this seeming contradiction.

Standard empirical cost studies explore how a firm's costs change as the firm changes scale or product mix, holding constant the number and other characteristics of rival firms. That approach fails to consider that the entry of new firms may impose higher interfirm networking costs on each incumbent or that the exit of incumbents (whether by merger or by failure) may reduce the interfirm networking costs to each surviving firm. Thus, when conventional studies suggest that a larger market should be supplied by additional firms rather than larger firms, they neglect the possibility that the increased networking costs resulting from entry may outweigh the higher unit costs resulting from increasing the size of a fixed number of firms. That is, in the presence of costly interfirm networking, the socially optimal structure may be more concentrated than the simple firm cost function would suggest and may require each firm to operate in a region of diseconomies of scale. This issue must be explored to establish valid public policy implications of empirical findings of diseconomies of scale in particular industries.[5]

This paper broadens the concept of cost structure to incorporate interfirm network effects, demonstrating how they alter the theoretical calculation of the socially optimal and unregulated equilibrium industry structures and presenting exploratory empirical evidence from a sample of U.S.

commercial banks. The empirical findings are consistent with the theory and suggest that hitherto overlooked networking costs may have important implications for public policy.

Before proceeding to a formal model, we note one property alluded to in the introduction. To expand its aggregate output, an industry must attain larger firms, more firms, or both. If interfirm networking is required and has a positive cost, the average cost will be an increasing function of the number of firms (as explored in the model below). Likewise, if firms have a U-shaped average cost structure, average cost is an increasing function of firm scale beyond some point. Thus, beyond some point, further expansion of an industry's aggregate output must drive up the average cost, whether that expansion is attained by expanding each firm or by entry. That is, *interfirm networking can undermine the neoclassical ability of an industry to maintain constant marginal cost at any aggregate scale* by means of entry of optimal-sized firms. This property is related to the decreasing returns to scale previously noted for certain efficient hierarchical networks by Radner (1992).

2. A Model of Payment System Networking Among Banks

To illustrate the properties of interest, we depict a simple model of a payment system network comprising n banks. As noted above, banks can or must interact in a variety of ways. Here we focus on just one of these elements to permit precise characterization of the process. Though derived for banks, the model may roughly characterize some other service industries also.

For the most part, we shall work with a symmetric structure in which each bank processes identical numbers of accounts and transactions, and in which transactions are uniformly distributed among accounts. This assumption of isotropic exchange parallels that in Laffont et al. (1996) and McAndrews and Roberds (1997). To isolate the network diseconomies per se, we posit a fixed aggregate number of accounts, M, and of net or "endpoint" transactions per unit of time, m. For example, each depositor writes a fixed number of checks per month, some on other accounts in the same bank (commonly called "on-us" transactions) and the rest on accounts in other banks ("transit" transactions), with the mix between on-us and transit transactions determined by the bank's market share measured by the number of accounts. We shall quantify how networking among banks unambiguously increases the total number of transactions by layering intermediate (i.e., transit) transactions onto the fixed base of m endpoint transactions.

The essential feature of the model is an explicit accounting of how m total endpoint transactions are apportioned and linked among the n banks. It will be necessary to distinguish between incoming transit and outgoing transit transactions, even though the banking industry reports only the latter of these components (e.g., in the Functional Cost Analysis survey administered by the Federal Reserve). A check deposited in bank A and drawn on an account in bank B will be considered as both an outgoing transit transaction with respect to bank A and an incoming transit transaction with respect to bank B. Here, "outgoing" and "incoming" refer to the direction in which interbank authorizations--not funds--flow. A distinctive feature of transit transactions is that they each show up at least twice, once at the outgoing bank and again at the incoming bank. If an intermediary such as a clearinghouse, the Federal Reserve, or a correspondent bank is used, such a transaction will move through at least three institutions.

The cost function analyzed here does not explicitly model the interbank market and is consistent with any given combination of direct interbank linkages and clearing houses (including endogenously developed combinations, with or without the check clearing role of the Federal Reserve). If banks have capacity constraints on their ability to deal directly with each other, a costminimizing structure will involve multiple vertical layers of interaction among banks, quantitatively exacerbating the network diseconomies analyzed below; we do not model this more extreme case.

Numbers of Transactions

In the symmetric structure with n banks, each bank holds a(n) = M/n accounts and directly encounters m/n transactions with the public each period, not including incoming transit transactions from other banks. Of these m/n transactions, a fraction 1/n will be on-us and the remainder will be outgoing transit, under the assumptions. The corresponding numbers of transactions per bank are b(n) = m/n^2 on-us and c(n) = $(n - 1)m/n^2$ outgoing transit. At the same time, the bank receives incoming transit transactions corresponding to a fraction 1/(n - 1) of each other bank's outgoing transit transactions. As there are (n - 1) other banks, the total number of incoming transit transactions for a given bank is therefore $(n - 1)m/n^2$, the same as the number of outgoing transit transactions. In aggregate, the total number of incoming transactions must equal the number of outgoing transactions; in the symmetric case, this equality must also hold for each bank individually. The total number of transit transactions per bank, including both incoming and outgoing, is $2(n - 1)m/n^2$. The number of all transactions, including both transit and on-us, is $(2n - 1)m/n^2$ per bank. In practice, the numbers of accounts and transactions will vary across banks and stochastically over time; we shall shortly relax the assumption of symmetry across banks, but will ignore stochastic variation for clarity of exposition.

Aggregated across all banks, the total number of transactions in the industry equals m/n on-us transactions plus 2m(1 - 1/n) transit transactions, or m(2 - 1/n) transactions in all. As the number of banks increases from n to n+1, total transactions increase by $m/(n^2 + n)$, which is positive for all n > 0. That is, not surprisingly, spreading a given number of endpoint transactions among a larger number of banks will entail additional intermediary (or transit) transactions and thereby *increase the total number of transactions processed* in the economy.

The implications of this elementary networking property for costs and optimal industry structure depend on details of the technology. In a neoclassical industry without networking, the socially optimal structure allows each firm to operate at its minimum average cost, in a region of locally constant returns to scale (except in the case of natural monopoly). With the networking effects characterized above, by contrast, it is straightforward to show that *the number of banks at which total industry cost is minimized can require each bank to operate in a region of diseconomies of scale,* whenever the cost of processing transit transactions is sufficiently large relative to the cost of processing on-us transactions. This property is explored in the next section.

In the (partially) asymmetric case in which firm i processes m_i (rather than m/n) endpoint transactions and each other firm $i \neq j$ processes m_j endpoint transactions (m_j = constant for all $i \neq j$), the number of on-us transactions processed by firm i is $m_i^2/[m_i + (n - 1)m_j]$. The number of incoming transit transactions is $m_i m_j(n - 1)/[m_i + (n - 1)m_j]$ and the number of outgoing transit transactions is the same. In this case, note that changing the number of accounts in bank i directly alters its number of on-us and outgoing transit transactions, but also indirectly changes its number of incoming transit transactions. The share of on-us transactions, as a fraction of all transactions, is $m_i/[m_i + 2(n - 1)m_j]$ for bank i. These results will be needed in the following sections.

Cost Structure

In the most general case, total costs will vary with the number of accounts as well as with the number of transactions. Each account entails such fixed costs as keeping records, filing statements, and monitoring account balances. In addition, we may suppose that the cost of processing an on-us transaction can differ from that of processing a transit transaction. The fact that banks have chosen not to report incoming transit transactions separately from outgoing transit transactions in their cost accounting suggests that there are no meaningful distinctions in the unit cost of processing the two types of transit transactions. In the symmetric case in which firms have identical but general cost functions, therefore, we can represent the total cost of bank i as:

$$(1) \qquad TC_i = C(a(n), b(n), c(n))$$

which is a monotone increasing function of each argument a(n), b(n), and c(n) as defined in the previous section.[6] Total industry cost is just $TC = nTC_i$ in the symmetric case. For given aggregate M and m, total industry cost is minimized by the value of n at which $0 = TC_i + n\partial TC_i/\partial n$, or:

$$(2) \qquad 0 = TC_i - (M/n)C_1(.) - (2m/n^2)C_2(.) + (2 - n)(m/n^2)C_3(.)$$

if the second-order condition holds (i.e., if the Hessian of C(.) is positive definite), where C_j denotes the first partial derivative of C(.) with respect to its jth argument.

Comparing this condition to the property of firm-level scale economies requires mild additional assumptions since the distinction between transactions and accounts essentially means that the bank is a multiproduct firm, even if it offers only a single type of account. We therefore adopt the neutral assumption that endpoint transactions are proportional to the number of accounts, or $m_i = \alpha M_i$ for each bank.[7] Since the number of accounts is a stock figure whereas transactions are a flow, it is possible to normalize the time interval under consideration such that $\alpha = 1$, and we adopt this convention.[8] Conventional scale economies are defined for bank i taking m_j as given. For the values of b(n) and c(n) given in the previous section, the scale elasticity of cost is:

$$(3) \qquad (m_i / TC_i)\ \partial TC_i/\partial m_i = m_iC_1/C + (m_iC_2/C)[m_i^2 + (n - 1)m_im_j] / [m_i + (n - 1)m_j]^2$$
$$+ (m_iC_3/C)m_j^2(n - 1)^2 / [m_i + (n - 1)m_j]^2$$

where the arguments of the bank's cost function C(.) have been suppressed for brevity. Expression (3) equals or exceeds 1 if there are constant or decreasing returns to scale, respectively, and is less than 1 if there are economies of scale.

To explore conditions under which the socially optimal industry structure entails diseconomies of scale at the firm level, we note that equation (2)--the condition for social optimality–-can be rewritten as:

$$(4) \qquad C = mC_1/n + 2mC_2/n^2 - (2 - n)mC_3/n^2$$

while the condition (3) > 1--reflecting diseconomies of scale--can be expressed in the symmetric case (where $m_i = m_j = m/n$) as:

$$(5) \qquad C < mC_1/n + mC_2/n^2 + (n - 1)^2mC_3/n^3.$$

Conditions (4) and (5) are simultaneously satisfied only if the right-hand side of (5) exceeds that of (4) or, equivalently, if:

$$(6) \qquad C_3 > nC_2.$$

Condition (6) is thus the necessary and sufficient condition in this framework for the socially optimal industry structure to entail diseconomies of scale at the firm level.[9] To explore the likelihood of this outcome, we first note that condition (6) holds for m sufficiently large relative to n, given $C_2 > 0$ and $C_3 > 0$. There are nearly 10,000 commercial banks in the U.S., besides several thousand other depository institutions. Because of localized markets, not every institution competes with all others. Nevertheless, the deconcentrated structure of U.S. banking, in conjunction with condition (6), may imply a *socially optimal industry structure that exhibits diseconomies of scale at the firm level* in U.S. banking even if C_3 is substantially smaller than C_2.

A major caveat is that this model abstracts from the costs of banks' other services (such as lending), which may not exhibit the same degree of networking. Moreover, the ratio m/n is smaller for some other networked services, such as large-payment electronic wire transfers, possibly leading to a violation of condition (6); such factors would tend to offset the conclusion of the previous paragraph or even cause firm-level economies of scale at the socially optimal industry structure.

Nevertheless, the policy implications of the foregoing calculations are striking. They imply that, unless other factors offset the pattern of networking costs among transaction accounts, we should expect that socially beneficial consolidation could occur in the U.S. banking industry, and perhaps in other large economies as well, beyond the point at which surviving banks are observed to be operating in a region of diseconomies of scale. Any offset would have to be exact if the classical result of locally constant returns to scale at the social optimum is to be preserved, an outcome that occurs on a set of measure zero. Previous empirical banking research has focused on the firm-level cost structure without taking account of network interactions, and has thereby overlooked a class of effects that the above analysis suggests should be significant.

An Example with Constant Marginal Costs

To characterize the quantitative and qualitative impact of interfirm networking on optimal industry structure more precisely, we must introduce additional assumptions about the bank's cost function. We shall consider an example with constant returns to scale to isolate the role of network diseconomies, in contrast to the situation explored in the previous section where diseconomies of scale could arise.

In the simplest case, assume that each on-us transaction has a constant marginal cost $k > 0$, each transit transaction (incoming or outgoing) has a constant marginal cost $K > 0$, and there are no other costs. This means, in particular, that the cost of a given number of incoming transit transactions is independent of the number of originating institutions, an assumption that is conservative because the marginal cost of an additional respondent bank is actually likely to be positive rather than zero, ceteris paribus. The total cost of each bank is $TC = km/n^2 + 2K(n - 1)m/n^2$, based on the numbers of transactions derived above, while the total social or aggregate cost is $nTC = km/n + 2K(n - 1)m/n$.

As the number of banks increases from n to $n + 1$, maintaining symmetry and holding m fixed, the aggregate total cost changes by $m(2K - k)/(n^2 + n)$, which takes the sign of $2K - k$. The aggregate cost is therefore lower in the less concentrated industry structure if and only if the marginal cost of an on-us transaction is more than twice as high as that of a single transit transaction. This result is intuitive because each transit transaction must be processed twice (once at each end), whereas each on-us transaction is processed only once. At the extremes of structural concentration, aggregate cost reduces to km for monopoly ($n = 1$, where all transactions are on-us) and approaches 2Km as $n \rightarrow \infty$ (where all transactions are transit). It is also apparent from these expressions that no interior extremum exists for aggregate costs in the linear case, but either an atomistic structure or monopoly will minimize social costs, depending on the relative magnitudes of K versus k.

Typically, we might expect to observe $k > K$, because a single transit transaction does not involve every part of a complete on-us transaction and should therefore be less costly.[10] We should also expect to observe $k < 2K$, since there is some duplication of steps (and therefore of costs) between an outgoing transit transaction and an incoming transit transaction, and because $k > 2K$ would imply that it is cheaper to split up every transaction between two banks than to route it through a single bank. Thus, to summarize, we should expect to find $K < k < 2K$ in practice. In this case, *the industry would be a natural monopoly from the standpoint of aggregate total costs, despite constant returns to scale at the individual bank level*--again an intuitive (though extreme) outcome since, when each combination of one outgoing and one incoming transit transaction is more costly than an on-us transaction, the monopoly structure minimizes aggregate costs by avoiding all transit transactions.

Entry Considerations

Determining the socially optimal industry structure is a separate issue than whether a free-entry equilibrium will attain that structure. In fact, within the linear framework of the previous section, it is easy to show that the monopoly structure is not sustainable against small-scale entry. Let bank i be a monopoly incumbent and bank j be an entrant. The incumbent's average cost is:

(7) $AC_i = (km_i + 2Km_j)/(m_i + 2m_j)$

and similarly for the entrant. The difference between the two average costs is:

(8) $AC_i - AC_j = 2(k - K)(m_i + m_j)(m_i - m_j)/[(m_i + 2m_j)(m_j + 2m_i)]$

which has the sign of $m_i - m_j$. Thus, if $m_i > m_j$, then $AC_i > AC_j$, so that a small-scale entrant has a cost advantage relative to the incumbent and will be able to underprice the incumbent and attract market share.

This example carries several salient implications. First, it demonstrates that *network diseconomies can have an asymmetric impact on the costs of individual banks*, if banks have different market shares. Second, it establishes that *excess entry can occur, even if it has the effect of driving up costs for all firms in the industry*. Note that the average cost of each bank can be an increasing function of the scale of either bank: $\partial AC_i/\partial m_i = 2m_j(k - K)/(m_i + 2m_j)^2 > 0$, since $k > K$, while $\partial AC_i/\partial m_j = [2m_i(K - k) + 4m_jK(m_i - 1)]/(m_i + 2m_j)^2$, which can take either sign but is positive for some ranges of parameter values (and similarly when i and j are interchanged). Third, the example illustrates that network diseconomies can make the scale of operations interdependent among banks in a way that violates the standard assumptions of empirical cost functions and duality theory, potentially biasing conventional cost estimates. Thus, the effect of network diseconomies is not only to drive a wedge between firm-level scale economies and optimal industry structure (as analyzed in the previous section) but *also to distort empirical firm-level cost functions themselves.*

An aspect of these properties that may have special relevance for the U.S. banking industry is that any firms not subject to the exogenous networking requirement but able to compete for certain of the services traditionally provided by banks could enter selected product markets and enjoy a lower cost by avoiding the networking. For example, as the economy grows, the unit cost of providing payments system services such as checking accounts may increase. If this cost cannot be fully passed on to the depositor, demand deposits will become more costly over time relative to alternative sources of funds. Eventually other firms that do not rely on core deposits, such as finance companies, may come to enjoy a cost advantage over traditional banks in the production of loans. Conversely, if the cost of networking is passed on to the depositors, alternative institutions that limit their participation in costly networking may be able to attract some depositors away from banks, with Merrill Lynch's sweep account providing an example.

3. Empirical Framework

The analysis in the previous sections strongly suggests that U.S. banks should exhibit observable network diseconomies. In this section we estimate a simple extension of a standard empirical cost function to explore whether the data are consistent with this phenomenon. The test focuses on metropolitan statistical areas (MSAs) as delineating local geographic banking markets, without attempting to capture additional networking effects involving broader regions. Previous research and current regulatory practice both use MSAs as a common proxy for banking markets (see for example Whitehead, 1980 and Jackson, 1992), so it is believed that the primary networking effects (especially among demand deposits) will occur within these regions. If a measurable network cost effect can be observed within a typical MSA, any additional network effects involving broader regions would only serve to strengthen the overall impact of networking on banks' costs. Within this framework, we shall measure the extent of networking alternately by the number of banks operating in the MSA and by the Herfindahl-Hirschman Index of concentration (HHI, the sum of squared market shares) of bank deposits within the MSA. The HHI reflects the degree of asymmetry in bank sizes within the market, unlike the simple number of banks.

Although the model in section 2 portrayed costs as a function of the numbers of accounts and transactions, the majority of recent empirical banking cost studies have used the dollar volume of various subsets of assets and liabilities as the measure of scale (see footnote 6). Because numbers of accounts and transactions are not reported for most banks, and in view of the strong positive relationship between numbers of accounts and transactions versus dollar volume, we adhere to the standard empirical practice in this section. We measure the scale of a bank as the outstanding dollar amounts of commercial and industrial loans, consumer loans, other loans, and securities and fed funds sold, as in the intermediation model.[11]

Although our extension controls for the number and relative scale of rival banks in each market, it maintains the conventional assumption that these factors are exogenous to each bank. The analysis of the previous section demonstrates that this is not strictly correct in the presence of network diseconomies, so the model fitted below should be regarded as a qualitative indicator of whether network effects may be present, rather than as a precise quantification of those effects. Nevertheless, measuring scale by dollars of assets rather than by numbers of transactions should minimize the distortion resulting from the interdependence between a bank's volume of transactions and the number of its local rivals.

The empirical model must control for agglomeration effects, given the focus of this study. Previous research (for example, Carlino, 1979; Moomaw, 1988; Calem and Carlino, 1991) has found that manufacturing firms' cost structures are affected by the size and characteristics of the communities in which they operate, for reasons other than network diseconomies, and it is reasonable to expect that similar effects may also occur in a service industry such as banking. For example, in a large city, economies of scale in complementary or support industries can drive costs down while congestion costs can have the opposite effect. If we did not control for such effects, they could possibly either mimic or else mask network diseconomies in our framework.

To control for agglomeration effects, each MSA's population and proportion of high school graduates were included as additional regressors in the model. In this framework, the quality of labor force education is measured as the percentage of the MSA's population above 18 years of age that graduated from high school. The present study, by adding network diseconomies to the recognized set of cost factors and applying the test to a firm-level sample of banking data, will also shed new light on previously unexplored agglomeration costs specific to the banking industry.

Sample

Our sample was drawn from all U.S. commercial banks operating exclusively within a single MSA as of year-end 1990, based on the location of branches accepting deposits.[12] Banks less than five years old were excluded from the sample because such banks typically have portfolio compositions and cost structures quite different from those of more mature banks. The year 1990 was chosen to coincide with the most recent available census population data for each bank's MSA. Previous research has found that the choice of year does not affect the estimates of scale economies in banking (Humphrey, 1990). Table 1 presents summary statistics on the data.

Banks with total assets less than $3 million or larger than $3 billion were excluded from the sample because of under-representation and for other reasons. Banks larger than $3 billion that maintain offices only within a single MSA are essentially wholesale in focus and are neither typical of the commercial banking industry nor subject to the same payment system networking phenomena that characterize a full-service, general-purpose bank. In addition, McAllister and McManus (1993) demonstrate that econometric problems can result from fitting a single specification across too large a range of scale (also implied in Barnett and Lee, 1985). Finally, White (1980, p. 154) has shown that a skewed sample distribution yields inconsistent OLS estimators, and truncating the two tails leaves a much less skewed sample distribution.

Because previous research has indicated that unit banks and branch banks face differing cost structures, and because the network aspects of unit banks and branch banks differ, we estimated separate cost functions for these two groups of banks.[13] The branch bank equation was fitted to 1971 banks while the unit bank equation was fitted to 877 banks.[14] Together, these two subsamples spanned 292 MSAs.

In any study of this type, spanning a large number of discrete markets, one must choose between fixed-effects and random-effects models in the treatment of the individual MSAs. Maddala (1987) has shown that, when the sample is not identical with the population of interest, inferences about true population values are more efficiently carried out by random-effects models than by fixed-effects models (see also Emmons, 1993, page 193). Since single-market banks operating in MSAs are a proper subset of the population of interest, and since many MSAs contain only one bank in each of our subsamples, we estimate only random-effects models.

The Estimating Equations

The functional form for each group was the standard translog, which has been the most widely used form for empirical cost studies in recent years. The translog cost function, augmented by agglomeration terms and proxies for network structure, is defined as:

$$\text{(9)}\qquad \ln C = \alpha_o + \sum_{i=1}^{4} \alpha_i Q_i + \sum_{j=1}^{3} \beta_j P_j + \tfrac{1}{2} \sum_{i=1}^{4} \sum_{k=1}^{4} \delta_{ik} Q_i Q_k + \tfrac{1}{2} \sum_{j=1}^{3} \sum_{k=1}^{3} \gamma_{jk} P_j P_k + \sum_{i=1}^{4} \sum_{j=1}^{3} \rho_{ij} Q_i P_j + \sum_{h=1}^{3} \eta_h X_h$$

where:

C = total costs (including both interest and noninterest expenses)
Q_i = $(\ln q_i - \ln {}_i)$ denotes the quantity of output i, where $\ln {}_i$ is the sample mean of $\ln q_i$;
Q_1 = commercial and industrial loans;
Q_2 = consumer loans;
Q_3 = other loans;
Q_4 = securities plus federal funds sold;
P_j = $(\ln p_j - \ln {}_j)$ denotes the price of input j, where $\ln {}_j$ is the sample mean of $\ln p_j$;
P_1 = price of labor, calculated as annual wage and benefit expenses per employee;
P_2 = price of funding, calculated as average interest costs per dollar of all deposits;
P_3 = price of physical capital, calculated as annual expenses on premises and equipment divided by the stock of these items;
X_1 = number of banks in MSA in one specification, and HHI in an alternate specification;
X_2 = population of MSA divided by 100,000;
X_3 = percentage of population above 18 years old graduated from high school;

We imposed the usual restrictions of symmetry and linear homogeneity in factor prices:

$$\delta_{ik} = \delta_{ki},\ \ \gamma_{jk} = \gamma_{kj},\ \ \Sigma\, \beta_j = 1,\ \ \Sigma_j\, \rho_{ij} = 0 \text{ for all i, and } \Sigma_j\, \gamma_{jk} = 0 \text{ for all k.}$$

Cost minimization implies the following factor share equations according to Shephard's lemma:

$$(10) \qquad \theta_j = \beta_j + \sum_k \gamma_{jk}P_k + \sum_i \rho_{ij}Q_i \qquad \text{for } j = 1 \text{ to } 3$$

where θ_j is the proportion of total cost spent on factor j. To avoid singularity, the capital share equation was omitted. The remaining share equations were jointly estimated with (9) to improve efficiency, using seemingly unrelated regression. We did not correct for possible heteroscedasticity, following standard practice in the empirical cost literature as well as Mishkin's (1990) analysis, which indicates that such adjustments can actually degrade statistical inference.[15]

Within this specification, the hypothesis of network diseconomies would predict a positive coefficient η_1 on the number of banks in the market, or a negative coefficient on the HHI. Urbanization economies would show up as a negative coefficient η_2 on population, although congestion costs could lead to the reverse finding. A negative coefficient η_3 on the high school graduation variable could reflect higher productivity of a better-educated workforce.

Results

Table 2 shows the estimated regression coefficients and t-statistics. The fit is good for each subsample, with adjusted R-squares ranging from 0.89 to 0.95. Most of the coefficients are statistically nonzero at conventional levels. The marginal cost of each output is positive in both regressions at sample mean values of input prices and output levels, as indicated by positive point estimates of α_i, consistent with monotonicity of the cost function. Likewise, predicted factor cost shares were positive at all sample values of the regressors in each regression, as required by monotonicity. A chi-square test rejected separability of the cost function at the 0.99 level.[16]

The coefficient on the number of banks, η_1, is significantly positive at the 0.01 level for both the branch bank and unit bank samples. Similarly, the coefficient on HHI is significantly negative at the 0.01 level for both samples. These results suggest that *a typical bank faces higher costs when its primary geographic market area contains a larger number of other banks*, consistent with the theory presented in the previous section. In the nomenclature of the literature on agglomeration economies, this effect corresponds to "localization diseconomies" where localization refers to the size of the industry within the community rather than the size of the community itself. The magnitude of the estimated effect appears modest at the individual bank level, with each additional bank increasing the cost of a branching bank incumbent by 3 basis points, or of a unit bank incumbent by 4 basis points. At the sample mean cost figures, the corresponding annual cost of an additional rival would be $3000 to a branching bank and $2000 to a unit bank.

These figures, while seemingly small, imply substantial aggregate networking costs. Comparing the implied cost of the sample mean market structure with that of an alternative (counterfactual) monopoly structure, we find that the average branching bank has costs that are higher by 2.24 percent (or nearly $229,000 per year) than in the monopoly structure, while the average unit bank has costs that are higher by 3.85 percent (or $201,000 per year) than in the monopoly structure.[17] Taking the lower dollar figure times the branch bank average of 75 banks per MSA implies an annual cost of bank networking of at least $15 million per MSA on average. Aggregated across all 292 MSAs in the sample, this figure implies nationwide networking costs of nearly $4.4 billion per year, or about 1.4 percent of the 1990 aggregate banking costs of $320 billion. Structural consolidation could reduce or eliminate these networking costs but would not necessarily result in an overall cost saving if there are diseconomies of scale at the firm level.

These estimates reflect only *local* bank presence and thus exclude any cost of networking with out-of-market banks. Actual networking costs should therefore exceed these estimates. Although not every bank will interact with every other bank across the country, and out-of-market interactions will typically be weaker than intramarket interactions, the sheer number of banks (more than 12,000 across the U.S. in the sample year) suggests that the omitted component of networking costs could also be substantial.[18]

Though alternative explanations besides interfirm network diseconomies could account for the observed sign of η_1, some explanations can be ruled out. For example, if banks in concentrated markets exert monopsony power, bank costs would be higher in markets containing many banks because wages and deposit interest rates would be higher in those markets. However, our regressions control for input prices so, whether or not our sample exhibits monopsony power, this hypothesis cannot explain the positive estimate of η_1. Another possible explanation involves agency costs in unconcentrated markets (Martin, 1993); our data cannot distinguish between this cause and network costs. In general, the implications of our results for network diseconomies should be regarded as suggestive rather than definitive.

Agglomeration Variables

Although the variables X_2 and X_3 are control variables unrelated to our hypothesis of network diseconomies, they are standard agglomeration variables with potential economic significance in their own right. The significantly negative coefficient on population indicates that banks face lower costs in larger cities, ceteris paribus, so that congestion effects appear to be more than offset by urbanization economies. One reason may be that a larger community permits information aggregation across borrowers that reduces a bank's unit costs of credit analysis. The estimated magnitude of this effect is twice as great for unit banks as for branch banks: increasing an MSA's population by 1 million reduces the cost of an average branch bank by a bit less than 2 percent and that of an average unit bank by about 3½ percent. Together with the positive estimate of η_1, this result suggests that localization effects and urbanization effects tend to go in opposite directions within the banking industry.

The high school graduation variable was negative and strongly significant for both samples, indicating that banks have lower costs in better-educated communities. One possible explanation could be that better educated workers are more productive for a given wage rate.[19]

Other properties of the cost function

Although it is not the primary focus of this paper to explore economies of scale or scope as conventionally measured, our estimates provide some information relevant to those issues. A chi-square test rejected constant returns to scale at the 0.99 level.[20] The scale elasticity of cost, defined as $SEC = \Sigma_i \, \partial \ln C / \partial \ln q_i$, is a common measure of ray scale economies. At the sample mean scale and input prices, its point estimate equals 0.957 for branch banks and 0.933 for unit banks, implying some economies of scale within this range for both samples. More generally, at sample mean input prices, the SEC was found to be:

(11) $SEC = 0.957 + 0.102\, Q_1 + 0.070\, Q_2 + 0.105\, Q_3 + 0.169\, Q_4$ and

(12) $SEC = 0.933 + 0.091\, Q_1 + 0.060\, Q_3 + 0.080\, Q_3 + 0.235\, Q_4$

for branch and unit banks, respectively. The estimates indicate statistically significant pairwise cost complementarities among all pairs of outputs in both samples, with the sole exception of consumer loans and other loans at unit banks, as implied by the δ_{ik} coefficients. Table 3 displays the output levels at which ray scale economies are exhausted for each output, calculated at sample mean values of input prices and of other output levels. These values are large compared with those found by most previous studies, especially with regard to consumer loans (see Berger and Humphrey, 1992b, for a survey), but a few studies have found similar or even unlimited minimum efficient scales (e.g., Shaffer, 1994). As noted above, our sample is less subject to econometric problems from skewness and a large range of scale than are many previous studies; to the extent that these differences contribute to the different findings, the results here may be more indicative of the actual cost structure. Other differences between our sample and previously studied samples may also contribute to the different estimates of efficient scale; for example, it may be that single-market banks have a larger efficient scale than multimarket banks. However, because our estimated minimum efficient scale lies beyond the upper bound of the sample range, these numbers represent an extrapolation that should not be regarded as definitive.

Although theoretical analysis in a previous section indicated that network diseconomies could distort conventional firm-level measures of scale economies where each firm's scale is erroneously interpreted as independent of the number of local rivals, this distortion did not appear to be large in the present sample. Replicating the regressions in Table 2 after omitting the number of banks in the same MSA yielded coefficient estimates that were essentially unchanged in sign and significance. Many of the coefficients, especially those critical to measures of scale or scope economies, were also nearly unchanged in magnitude. The scale elasticity of cost in these alternate regressions was estimated as 0.957 and 0.934, respectively, for the branch bank and unit bank samples at the sample mean values of the input prices. As noted above, the use of dollar-based output measures is likely to be a major factor in explaining the small magnitude of this distortion, given that the number of local rivals will mainly affect the number of a bank's payment transactions rather than the size of its loan or deposit base.

4. Conclusion

This paper has introduced the concept of network diseconomies associated with the intrinsic need for multiple firms within an industry to interface with each other in the provision of the industry's primary service. Several industries were discussed as possible examples, and a formal theoretical model characterized transaction accounts as a source of network diseconomies in the banking industry. Further analysis established conditions under which the socially optimal industry structure is more concentrated than that corresponding to either the minimum average cost point for each firm (locally constant returns to scale) or the free-entry equilibrium; a rough calibration of the model suggests that some of these conditions may hold for the U.S. banking industry. An additional example particularly relevant to the payments system demonstrated that interfirm linkages can result in a natural monopoly despite constant returns to scale at the firm level as conventionally measured. These various results demonstrate that the concept of network diseconomies can have important public policy implications.

In addition, some simple empirical tests were presented to explore in a general way whether data from the U.S. banking industry are consistent with the possibility of network diseconomies. The empirical model also incorporated a test for agglomeration effects, which appears to be the first application of such a test to the banking industry. Evidence consistent with both network diseconomies and agglomeration effects was found, with branch banks and unit banks exhibiting localization diseconomies and urbanization economies.

The potential relevance of network diseconomies to understanding structural trends and informing public policy may be increasing due to recent or ongoing developments in several industries. Within the U.S., competition in the telecommunication industry and deregulation in the airline industry is only a few years old, with lessons still being learned. Endogenous structural consolidation in the banking industry, in Europe and the U.S., often appears at odds with the standard interpretations of conventional cost studies (Berger and Humphrey, 1992b). The continued erosion of geographic market barriers and the globalization of many industries, including banking, is imposing the need for a widening sphere of interfirm networking in the normal course of business.

To the extent that the costs of increased networking offset the widely recognized benefits of broader market integration, the effects identified in this paper may suggest an upper limit to the beneficial degree of globalization. Previous studies have identified other impediments to integration, including costs associated with potential mismanagement of assets (Williamson, 1985), influence costs (Milgrom and Roberts, 1990), and other agency costs (Olsen, 1996), but these studies have focused on intrafirm integration. The effects identified in this paper both add to the list of recognized barriers to integration and extend the research on such barriers to interfirm and market integration. Moreover, since broader integration of markets in any industry will generally necessitate corresponding linkages in the payments and communication systems, a complete accounting of interfirm networking costs must include these factors even where they are ancillary to the industry in question. In these and similar instances, further research on the effects of network diseconomies may be needed both to quantify the problems and to identify solutions.

Table 1

Summary Statistics

Variable	**Branch Banks**		**Unit Banks**	
	Mean	**Standard Deviation**	**Mean**	**Standard Deviation**
Log (total cost)	9.232	0.888	8.561	0.772
Commercial loans	35841	56648	18756	37219
Consumer loans	20791	47763	10587	38339
Other loans	54540	92087	22974	61943
Securities and fed funds sold	55857	89778	29307	53539
Wage rate	27.93	6.02	29.41	7.16
Deposit interest rate	0.072	0.007	0.071	0.007
Price of physical capital	0.373	0.287	0.489	0.501
Number of banks in MSA	75.1	85.4	95.5	90.2
Population of MSA (100,000)	17.87	20.69	18.48	21.17
Percentage of high school graduates	0.774	0.062	0.787	0.065

(Output levels are in $000.)

Table 2
Regression Results

	Branch Banks				Unit Banks			
Coeff.	**Est.**	**(t-stat.)**	**Est.**	**(t-stat.)**	**Est.**	**(t-stat.)**	**Est.**	**(t-stat.)**
α_o	9.995	(213.41*)	10.015	(209.22*)	9.278	(98.87*)	9.316	(95.77*)
α_1	0.197	(30.65*)	0.196	(30.49*)	0.235	(19.83*)	0.233	(19.75*)
α_2	0.159	(28.62*)	0.161	(28.80*)	0.177	(19.67*)	0.179	(19.85*)
α_3	0.321	(47.61*)	0.317	(47.46*)	0.265	(21.87*)	0.263	(21.95*)
α_4	0.280	(36.53*)	0.283	(37.10*)	0.256	(16.80*)	0.261	(17.46*)
β_1	0.204	(115.85*)	0.204	(115.97*)	0.215	(59.74*)	0.214	(59.79*)
β_2	0.532	(185.50*)	0.532	(185.51*)	0.525	(87.73*)	0.525	(87.68*)
δ_{11}	0.102	(16.14*)	0.101	(15.98*)	0.091	(8.84*)	0.088	(8.68*)
δ_{12}	-0.019	(-4.71*)	-0.019	(-4.57*)	-0.015	(-2.55*)	-0.015	(-2.54**)
δ_{13}	-0.042	(-8.98*)	-0.042	(-9.03*)	-0.028	(-3.92*)	-0.029	(-4.03*)
δ_{14}	-0.060	(-9.40*)	-0.059	(-9.28*)	-0.071	(-6.14*)	-0.068	(-5.97*)
δ_{22}	0.070	(19.23*)	0.070	(19.20*)	0.060	(11.48*)	0.060	(11.53*)
δ_{23}	-0.016	(-4.44*)	-0.016	(-4.41*)	0.013	(2.10**)	0.013	(2.20**)
δ_{24}	-0.036	(-7.03*)	-0.035	(-6.98*)	-0.049	(-5.50*)	-0.049	(-5.53*)
δ_{33}	0.105	(25.24*)	0.103	(24.89*)	0.080	(13.18*)	0.080	(13.18*)
δ_{34}	-0.040	(-5.72*)	-0.038	(-5.54*)	-0.054	(-4.85*)	-0.054	(-4.91*)
δ_{44}	0.169	(16.24*)	0.167	(16.08*)	0.235	(10.86*)	0.233	(10.82*)
ρ_{11}	-0.002	(-1.11)	-0.002	(-1.11)	0.007	(2.45**)	0.007	(2.47**)
ρ_{12}	0.002	(0.70)	0.002	(0.75)	-0.010	(-2.23**)	-0.010	(-2.20**)
ρ_{21}	0.005	(3.41*)	0.005	(3.38*)	0.005	(2.24**)	0.005	(2.25**)
ρ_{22}	-0.017	(-7.86*)	-0.017	(-7.98*)	-0.018	(-5.36*)	-0.018	(-5.35*)
ρ_{31}	-0.006	(-3.69*)	-0.006	(-3.68*)	-0.006	(-2.04**)	-0.005	(-1.98**)
ρ_{32}	0.008	(3.30*)	0.008	(3.34*)	0.010	(2.40**)	0.011	(2.51**)
ρ_{41}	-0.012	(-5.81*)	-0.012	(-5.82*)	-0.024	(-6.37*)	-0.024	(-6.48*)
ρ_{42}	0.018	(5.63*)	0.018	(5.65*)	0.025	(4.40*)	0.024	(4.26*)
γ_{12}	-0.141	(-19.86*)	-0.142	(-20.01*)	-0.119	(-10.02*)	-0.119	(-10.06*)
γ_{13}	0.027	(11.51*)	0.027	(11.52*)	0.016	(4.82*)	0.016	(4.80*)
γ_{23}	-0.055	(-14.94*)	-0.054	(-14.92*)	-0.051	(-9.69*)	-0.051	(-9.70*)
η_1=# banks	.00030	(3.90*)	--		.00040	(2.59*)		--
η_1 = HHI		--	-0.245	(-4.15*)		--	-0.376	(-2.90*)
η_2	-.00187	(-5.79*)	-.00116	(-5.91*)	-.00348	(-5.03*)	-.00240	(-6.22*)
η_3	-0.322	(-5.38*)	-0291	(-4.98*)	-0.520	(-4.40*)	-0.486	(-4.30*)
n	1971		1971		877		877	
R-adj.	0.95		0.95		0.89		0.89	

*significant at 0.01 level. ** significant at 0.05 level.

Table 3

Sizes at which Ray Scale Economies Are Exhausted for Each Output (calculated at sample mean values of input prices and other output levels)

	Branch Banks		Unit Banks	
Output	**Multiple of Sample Mean**	**$ Billions**	**Multiple of Sample Mean**	**$ Billions**
Commercial Loans	2564	91.9	4362	81.8
Consumer Loans	162,000	3368	918,900	9728
Other Loans	646.1	35.2	10,115	232.4
Securities and fed funds sold	71.07	4.0	23.64	0.7

References

Barnett, William A. and Yule W. Lee, 1985, "The Global Properties of the Minflex Laurent, Generalized Leontief, and the Translog Flexible Forms," *Econometrica* 53, 1421-1437.

Berger, Allen N., Gerald A. Hanweck, and David B. Humphrey, 1987, "Competitive Viability in Banking: Scale, Scope, and Product Mix Economies," *Journal of Monetary Economics* 20, 501-520.

Berger, Allen N. and David B. Humphrey, 1992a, "Measurement and Efficiency Issues in Commercial Banking," in Zvi Griliches, ed., *Output Measurement in the Services Sector*, NBER (Chicago: University of Chicago Press).

Berger, Allen N. and David B. Humphrey, 1992b, "Megamergers in Banking and the Use of Cost Efficiency as an Antitrust Defense," *Antitrust Bulletin* 37, Fall, 541-600.

Broecker, T., 1990, "Credit-Worthiness Tests and Interbank Competition," *Econometrica* 58 (2), 429-452.

Calem, Paul S. and Gerald A. Carlino, 1991, "Urban Agglomeration Economies in the Presence of Technical Change," *Journal of Urban Economics* 29 (1), January, 82-95.

Carlino, Gerald A., 1979, "Increasing Returns to Scale in Metropolitan Manufacturing," *Journal of Regional Science* 19, 343-351.

Caves, W. C., Laurits R. Christensen, and M. W. Tretheway, 1984, "Economies of Density Versus Economies of Scale: Why Trunk and Local Service Airline Costs Differ," *Rand Journal of Economics* 15, 471-489.

Economides, Nicholas, 1996, "The Economics of Networks," *International Journal of Industrial Organization* 14 (6), October, 673-699.

Emmons, William, 1993, "Increased Risk-Taking versus Local Economic Conditions as Causes of Bank Failures," *Proceedings of a Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, 189-209.

Farrell, Joe and Garth Saloner, 1985, "Standardization, Compatibility, and Innovation," *Rand Journal of Economics* 16, 70-83.

Ferrier, Gary D. and C. A. Knox Lovell, 1990, "Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence," *Journal of Econometrics* 46, 229-245.

Gehrig, Thomas, 1996, "Natural Oligopoly and Customer Networks in Intermediated Markets," *International Journal of Industrial Organization* 14 (1), 101-118.
Gilligan, Thomas, Michael Smirlock, and William Marshall, 1984, "Scale and Scope Economies in the Multi-Product Banking Firm," *Journal of Monetary Economics* 13 (3), May, 393-405.

Hancock, Diana, 1985, "The Financial Firm: Production with Monetary and Non-monetary Goods," *Journal of Political Economy* 93 (5), October, 859-880.

Humphrey, David B., 1990, "Why Do Estimates of Bank Scale Economies Differ?" Federal Reserve Bank of Richmond *Economic Review* 76 (5), September/October, 38-50.

Humphrey, David B., 1992, "Flow Versus Stock Indicators of Banking Output: Effects on Productivity and Scale Economy Measurement," *Journal of Financial Services Research* 6(2), August, 115-136.

Jackson, William E., 1992, "Is the Market Well Defined in Bank Merger and Acquisition Analysis?" *The Review of Economics and Statistics* 74, November, 655-661.

Katz, M. and Carl Shapiro, 1985, "Network Externalities, Competition, and Compatibility," *American Economic Review* 75, 424-440.

Katz, M. and Carl Shapiro, 1986, "Technology Adoption in the Presence of Network Externalities," *Journal of Political Economy* 94, 822-841.

Kim, Moshe and Uri Ben-Zion, 1989, "The Structure of Technology in a Multioutput Branch Banking Firm," *Journal of Business and Economic Statistics* 7 (4), October, 489-496.

Laffont, Jean-Jacques, Patrick Rey, and Jean Tirole, 1996, “Network Competition: I. Overview and Nondiscriminatory Pricing,” IDEI, Toulouse, mimeo.

Maddala, G.S., 1987, "Limited Dependent Variable Models Using Panel Data," *Journal of Human Resources* 22, 311-338.

Martin, Stephen, 1993, "Endogenous Firm Efficiency in a Cournot Principal-Agent Model," *Journal of Economic Theory* 59, 445-450.

McAllister, Patrick H. and Douglas A. McManus, 1993, "Resolving the Scale Efficiency Puzzle in Banking," *Journal of Banking and Finance* 17 (2-3), April, 389-405.

McAndrews, James and William Roberds, 1997, "A Model of Check Exchange," Federal Reserve Bank of Philadelphia Working Paper No. 97-16, Revised October.

Milgrom, Paul R. and John Roberts, 1990, "Bargaining and Influence Costs and the Organization of Economic Activity," in J. E. Alt and K. A. Shepsle, eds., *Perspectives on Positive Political Economy* (Cambridge, MA: Cambridge University Press).

Mishkin, Frederick S., 1990, "Does Correcting for Heteroscedasticity Help?" NBER Technical Working Paper 88 (Cambridge, MA).

Moomaw, Ronald, 1988, "Agglomeration Economies: Localization or Urbanization?" *Urban Studies* 25 (2), April, 150-161.

Nakamura, Leonard I., 1993, "Loan Screening Within and Outside of Customer Relationships," Federal Reserve Bank of Philadelphia Working Paper No. 93-15.

Neave, Edwin H., 1991, *The Economic Organisation of a Financial System* (New York: Routledge).

Olsen, Trond E., 1996, "Agency Costs and the Limits of Integration," *Rand Journal of Economics* 27 (3), Autumn, 479-501.

Radner, Roy, 1992, “Hierarchy: The Economics of Managing,” *Journal of Economic Literature* 30 (3), September, 1382-1415.

Sealey, Calvin, Jr. and James T. Lindley, 1977, "Inputs, Outputs, and a Theory of Production and Cost at Depository Financial Institutions," *Journal of Finance* 32, 1251-1266.

Shaffer, Sherrill, 1994, "A Revenue-Restricted Cost Study of 100 Large Banks," *Applied Financial Economics* 4, 193-205.

White, Hal, 1980, "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review* 21 (1), 149-170.

Whitehead, David D., 1980, "Relevant Geographic Banking Markets: How Should They Be Defined?" Federal Reserve Bank of Atlanta *Economic Review*, January/February, 20-28.

Williamson, Oliver E., 1975, *Markets and Hierarchies* (New York: Free Press).

Williamson, Oliver, 1985, *The Economic Institutions of Capitalism* (New York: Free Press).

Footnotes

1. Gehrig (1996) analyzes a form of networking in an intermediation industry with consumer search costs, showing that investment by firms in information networks for marketing purposes can reduce the equilibrium degree of industry concentration. Although banking is a clear example of an intermediation industry, Gehrig's model has a completely different focus from ours and consequently reaches contrasting conclusions. Gehrig models networks as linking firms to consumers rather than firms to firms or consumers to consumers, analyzes consumer matching rather than costs, and construes the size of the network as a quality attribute rather than as an exogenous market characteristic.

2. Some industries--such as telecommunications, electricity, gas, and water--have become known as "network industries" because of their universal service mandate. There is an imperfect overlap between these industries and those under consideration in this paper, the major distinction being that public utilities such as electricity, gas, and water do not intrinsically link consumers together, but only provide a given service to each consumer. In such cases, interfirm networking would not be required even if the industry were not structured as a local monopoly.

3. Of course, the postal pricing structure could be altered to discourage the large volume of commercial advertising, but the growth in the aggregate demand *function* for postal services has been exogenous.

4. The total number of assisted and unassisted mergers from 1980 through 1994 was 7103, according to the FDIC's *Statistics on Banking, 1934-1994* (page A-10). By the end of 1994 there were 10,451 commercial banks in the U.S. (ibid.).

5. At least two previous empirical studies have taken account of certain influences of networks on the cost structure. Caves et al. (1984) distinguish between economies of scale and economies of density pertaining to airline routes. Kim and Ben-Zion (1989) apply the same concept to banking costs, reflecting the branching network of an individual bank. The present study is distinguished from these in its focus on the linkages among firms rather than those within a firm.

6. Focusing on the cost of accounts and transactions corresponds to the so-called "production model" of a banking firm in which output is measured as the number of accounts. This approach has been used in some empirical banking cost studies such as Gilligan et al. (1984) and Ferrier and Lovell (1990). The alternative "intermediation model" (Sealey and Lindley, 1977), along with its variants such as the user-cost model (Hancock, 1985) and the value-added model (Berger and

Humphrey, 1992a), has seen more use in recent empirical banking cost studies. These latter models measure scale in terms of the dollar volume of various subsets of assets and liabilities (which is reported for all banks), rather than the number of accounts (which is not reported for most banks). There is a strong positive relationship between the number of accounts and the dollar volume.

7. If anything, it seems reasonable to expect the number of transactions per account to increase with the number of other accounts in the economy. Otherwise, the implicit assumption is that, as the number of accounts in the economy grows, original accounts do not interact at all with new accounts. Thus, the assumption of proportionality is conservative with respect to network diseconomies.

8. Previous research has also noted the distinction between stocks and flows in the various measures of bank output; see for example Humphrey (1992).

9. Conversely, the socially optimal structure is characterized by locally constant returns to scale at the firm level as conventionally measured, if and only if the left-hand side of expression (6) equals the right-hand side, a condition that is satisfied only on a set of measure zero. Thus, the socially optimal structure nearly always entails some deviation from the minimum-average-cost firm scale when there is interfirm networking.

10. McAndrews and Roberds maintain the assumption that the cost of an on-us transaction equals that of a transit transaction, but note that this is not the most realistic case.

11. Alternative versions of the model were also estimated corresponding to the value-added model and the user-cost model, yielding results that were qualitatively and quantitatively similar.

12. MSAs in New England typically do not follow county boundaries and were replaced by New England County Metro Areas (NECMAs) in this study.

13. Kim and Ben-Zion (1989) define a bank's branches as a component of scale economies, requiring that the number of branches be included as a regressor. While this approach was appropriate for the focus of their study, it is not standard practice in the empirical banking literature because of equilibrium expansion path arguments (see Berger et al., 1987), and is

therefore not included in our model.

14. Some banks did not report a sufficient level of detail to be included in the model as specified, and these banks were excluded from the sample.

15. The Pearson correlation coefficient between cost residuals and total bank assets was 0.025 for the unit bank sample (not significantly different from zero) and 0.049 for the branch bank sample (significantly different from zero at the 0.03 level but still small in magnitude). These figures suggest that any monotonic pattern of heteroscedasticity in our sample is minimal.

16. The test statistics were 124 for branch banks and 277 for unit banks; the 0.99 critical value for chi-square with 8 degrees of freedom is 20.1.

17. The calculation is [exp(9.232 + (75 x 0.0002994)) / exp(9.232 + (1 x 0.0002994))] = 1.0224 for branching banks, and similarly for unit banks, where the mean number of banks per MSA is 75 for branching banks and 95 for unit banks.

18. Testing this broader aspect of networking costs would require a pooled time series / cross-sectional sample to obtain variation in the number of banks nationwide. The availability of relevant Census data only at 10-year intervals precludes the practical construction of such a test.

19. Another possible factor might be that better educated borrowers have lower delinquency and default rates, requiring less costly credit monitoring by the bank and imposing lower loan chargeoffs. We did not attempt to distinguish empirically among these potential explanations.

20. The test statistics were 657 for branch banks and 2042 for unit banks; the 0.99 critical value of the chi-square with 11 degrees of freedom is 24.7.