Financial Communities

Internet information flows as a basis for portfolio strategy.

Sanjiv R. Das and Jacob Sisk

Every asset manager is concerned with how groups of stocks interact with each other—this is the essence of portfolio management. Portfolios may be described in many ways, such as diversified, leveraged, or targeted, and classified by industry sector, product, currency, security type, or objective. In each case, classification serves a purpose; it summarizes the relationship among the stocks in the portfolio. We propose a new approach to grouping stocks, into financial communities based on small investor sociology.

We define a financial community as a group of stocks discussed by common responders. By analyzing a large database of web opinions expressed by millions of individuals, we uncover the structure of connections between and among stocks, and consider the implications of our findings for portfolio managers.

The popularity of the Internet as a medium of stock market discussion lets us examine the social structure of the information flow driving stock prices. We conclude that:

1. Graph-theoretic techniques may be used to describe the structure of financial communities. This provides an alternative metric for portfolio grouping. That there are such communities has a variety of implications for portfolio managers.
2. Stocks in tight communities display more *connectedness* of information flow than stocks in loose communities. Higher connectedness is shown to translate into higher return correlations,

Sanjiv R. Das is a _____ in the Santa Clara University Leavey School of Business in Santa Clara, CA.

Jacob Sisk is a ____ at Overture/Yahoo! Matching Sciences Research in Pasadena, CA.

suggesting that portfolios formed across communities offer more diversification than portfolios within communities.

3. Highly connected stocks provide better risk-return performance than less connected stocks, which predicates tilting portfolios toward those stocks.
4. Eigenvector techniques may be used to detect stocks that are hubs for information flow, using a sociological measure known as *centrality*. Stocks with high centrality scores tend to have greater average covariance with other stocks than those with low scores. This suggests that portfolio managers focus more effort on tracking central stocks than others.

RELATED LITERATURE

Our work is related to the literature on market microstructure, portfolio theory, and information generation on the web. The flow of information into prices is a central feature of a smoothly functioning market that incorporates collective opinion on firm fortunes into stock prices through trading (described in classic microstructure models; see Kyle [1985]). Exchange trading is a social activity as well as an economic one (see Baker [1984]). More recent research indicates that comovement of stocks may arise out of noise trader patterns, as in Barberis and Shleifer [2003] and Cornell [2004]; the related information flow may be detectable in investor discussion.

We examine how a common information flow via discussion on Internet stock message boards is related to stock returns across groups of stocks. We are interested not in the trading patterns of investors in a single stock, but rather in the information flow among groups of stocks by cohesive investors who coalesce into groups called *financial communities*. Using hundreds of message boards as empirical input (representing more than 23 million messages), and graph-theoretic tools as a modeling structure, we find that common information flow in financial communities is related to portfolio risk and return.

These stock message boards make the process of opinion formation quite observable. While many financial theorists assume a strong link between information and stock returns, the mechanics of this link have been relatively unexplored. Research on information cascades suggests we may be underestimating the role of community behavior. For a vivid survey, see Bikhchandani, Hirshleifer, and Welch [1996]. For details, see Welch [1992] and Watts [2002].

Researchers have begun to examine how web-based opinions and discussion relate to trading. Wysocki [1999] finds that overnight message posting volume is predictive of next-day stock trading volume and returns. Bagnoli, Beneish, and Watts [1999] provide evidence that "whisper" earnings forecasts posted to the web may be more accurate than those of First Call analysts. Antweiler and Frank [2004] show that messages weakly predict volatility but not returns, and Harris and Raviv [1993] that agreement among posted messages is associated with reduced trading volume. Das and Chen [2001] find that returns drive message board sentiment; the results are stronger when messages for many stocks are aggregated into an index-level sentiment measure.

Results suggesting that web discussion is not predictably related to returns are developed in Tumarkin and Whitelaw [2001], although Antweiler and Frank [2002] show that high message volume is usually a forerunner of high volatility and low returns. Das, Martinez-Jerez, and Tufano [2001], coining the term *e-information*, find that disagreement about market information prompts extensive debate, and message posting is a catalyst for impounding information into stock prices.

All this research suggests an examination of the community structure of web information might be informative for portfolio managers. Stock returns are known to be subject to specific cross-covariation effects, presumably emanating from common return factors, but also from commonality of the opinion formation process (see Boudoukh, Richardson, and Whitelaw [1994]). An analysis of general web discussion about the stock market seems warranted, because these group dynamics may be antecedents of market phenomena like herding, crashes, or bubbles.

There are many ways community structure translates into activities that relate to portfolio returns. Message boards are forums for people to seek validation of their opinions, both before and after making trades in a stock. Message posting is costless, and therefore occurs in high volume. Postings provide considerable information, but may also lead to multiple equilibria of differing informativeness (see Admati and Pfleiderer [2001]).

Possible behavior and transaction cost reasons may explain why message boards provide information affecting the correlation of stock returns. Barber and Odean [2002] show that individual investors tend to demonstrate "attention-based buying"—they buy more into stocks that trade more, return more, and generate more news. The greater the commonality of information across boards, the more likely it is that attention-based trading will drive returns closer together for stocks with high attention value. Message board discussion will reflect enhanced attention, which may be an underlying reason for some of our results. It may also be interpreted as "simplistic group think" as discussed in Zuckerman and Rao [2003].

We represent the financial community of stock message boards as a network graph. Stock tickers are nodes on the graph, and we develop metrics for the connection strengths between nodes on the graph. The graph model allows us to analyze specific questions about the nature of this community, its degree of *connectedness*, and how differences in connectedness over time covary with stock market returns.

This graph of message boards also enables the notion of the *centrality* of some stocks relative to others (see DeMarzo, Vayanos, and Zwiebel [2003] for a formal theoretical analysis of markets in general). The centrality property recognizes that some stocks gain more importance than others in the collective processing of information, and that their premier position on the information graph may affect how quickly information is impounded into stock returns.

We compute empirical properties on both connectedness and centrality from our information graph, and use these measures to analyze the portfolio implications of activity in financial communities.

NETWORK MODEL

We model the universe of stocks as a network, with nodes and connections between nodes (a formal description is provided in the appendix). Each node on the graph corresponds to one stock. The connection strength between nodes is determined by the number of common message posters between any two stocks in a given time, which we set at one month. If Alice and Bob, for example, post messages to the boards of IBM and Google in the same month, the connection strength between those two nodes would be 2 (assuming no other common posters).

This network graph may be summarized in an *adjacency matrix* A. This is a square matrix including as many rows and columns as stocks (nodes). This adjacency matrix is the basis for our analysis. The entry in cell $A[v, w]$ of the matrix is equal to the connection strength between nodes v and w. (Of course, $A[v, w] = A[w, v]$, as the matrix is symmetric.)[1]

DATA

Our data constitute all messages posted to stock boards from January 2000 through April 2001, representing a total of 16 months. Various provider boards are covered, but the major share of message volume comes from Yahoo's message boards. Other boards with material message volume are Motley Fool, Raging Bull, and Silicon Investor.

There are over 23 million posted messages across all boards. Screen names across all message boards totaled over 50,000, which represents a lot of people involved in the process of opinion formation. These statistics suggest that message boards are now an accepted and active medium of information exchange for the equity markets.

The total number of tickers covered in this study is over 2,000. The data come from Codexa, Inc., a firm that harvests all Internet information traffic on U.S. stocks. A major part of the information constitutes postings to message boards, the focus of our study.

We extract and store from each message the ticker, the poster's screen name, and the full time stamp. From this, we are able to compute the adjacency table $A(v, w)$ listing the number of posters in common across each pair of tickers (v, w) within each month. We use the matrix to develop various measures of the information linkages between tickers.

CONNECTEDNESS

We define nodes v and w as connected if there is a path of non-zero edge weights between nodes in a chain:

$$\{v \rightarrow v_1 \rightarrow v_2 \rightarrow \ldots \rightarrow v_k \rightarrow w\}$$

where $k = 0$ is possible as well.

The questions we ask relate to the degree of connectedness of the graph, such as the number of connections and the strength of these connections; a community is thus characterized by the number and strength of its relationships. The nodes fall into groups of connected stocks. Some stocks are orphaned and have no community affiliation (we call these *singletons*).

By means of simple graph algorithms, we determine how many communities there are each month, as well as their size.[2] We identify communities for quantitative and qualitative reasons: quantitative because we want to determine the number and size of communities in any month, and qualitative so we can examine the connection between community structure and stock returns.

Thresholds

Connectedness is related to the extent of overlapping posters across message boards. Hence, if board A has common posters with board *B*, and board *B* has common posters with board *C* (different from those with board *A*), then *A* is also connected to *C*, and the three boards together represent a community. It turns out that most message boards are connected if we require only one common poster across any two boards.[3]

The presence of common posters results in opinion flow among connected boards. The strength of information flow depends on the number of common posters. To incorporate this feature, we specialize our definition of connectedness with respect to an *overlap threshold*, denoted *K*, i.e., two boards are connected only if they have at least *K* common posters, $A(v, w) \geq K$. Setting $K = 1$ (minimal level) results in one large community and many singleton communities. As *K* increases from 1, though, we get interesting community patterns.

Exhibit 1 is an example of a connectedness diagram. There are seven tickers, *A* through *G*, represented by the nodes on the graph. The numbers represent the number of common posters across the tickers' message boards. In this graph, the connection threshold is set to $K = 1$, meaning that two tickers are connected if there is at least one common poster on their message boards. Hence, all connections are valid, whatever their strength. In this example, we can see that all the stocks form a single community, as they are all linked.

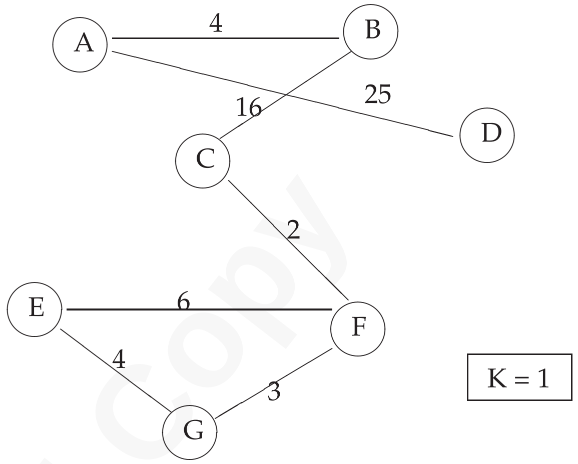
In Exhibit 2 we set the connection threshold to $K = 5$, and the result is fewer connections. Instead of one large community, we get four communities: $\{A, D\}$, $\{B, C\}$, $\{E, F\}$, and $\{G\}$. Thus, there are three small communities, and one singleton community.

Community Structure

Defining communities according to different overlap thresholds has two effects. First, as *K* increases, the size of the biggest community drops, and instead of one single large community, we expect to see smaller but more numerous communities.

Second, as *K*, the threshold number of common posters increases, we also expect to see fewer eligible message boards, as some boards may have fewer than *K* posters, and we eliminate all boards that do before segregating firms into communities. We thus ensure that we choose boards with a certain minimum amount of information flow. As the number of eligible message boards declines, there are also fewer communities (and we get more singletons).

E X H I B I T 1
Connectedness Graph—*K* = 1



The interaction of the two effects suggests that as *K* increases, large communities become smaller and devolve into a few smaller ones, and then as *K* gets much larger, the total number of communities drops.

We analyze messages in monthly blocks. Hence, for each calendar month over January 2000–April 2001, we compute the number of communities for a range of threshold levels, where *K* takes values in the set: $\{1, 2, 5, 10, 25, 50, 100, 200\}$. For a visual representation of the connected community, we use a spatial graph algorithm to draw the network graph for any threshold level *K*.

Exhibit 3 presents one such graph for one month (February 2001). The threshold level is $K = 25$, and nodes with fewer than 25 posters are eliminated before determining communities. The figures for the other months show similar graph structure.

Exhibit 4 presents a detailed breakdown of the community structure. For each month, two numbers are reported for each threshold (*K*) size: 1) The first number is the number of singletons, and 2) the second number is the size of the largest community. As an example, consider January 2000, at the overlap threshold $K = 25$. There are 392 singletons, and the largest community is composed of 104 stocks.

At a threshold of $K = 1$, almost all stocks are connected to each other. Hence, community structure tends to represent one huge community and many singletons. (In the

Exhibit 2

Connectedness Graph with Higher Threshold—*K* = 5

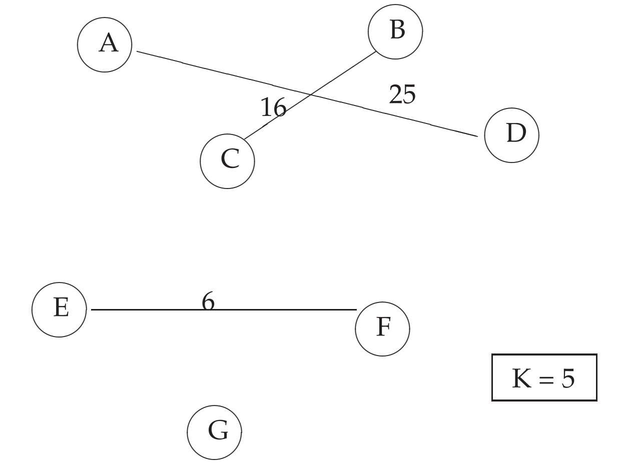
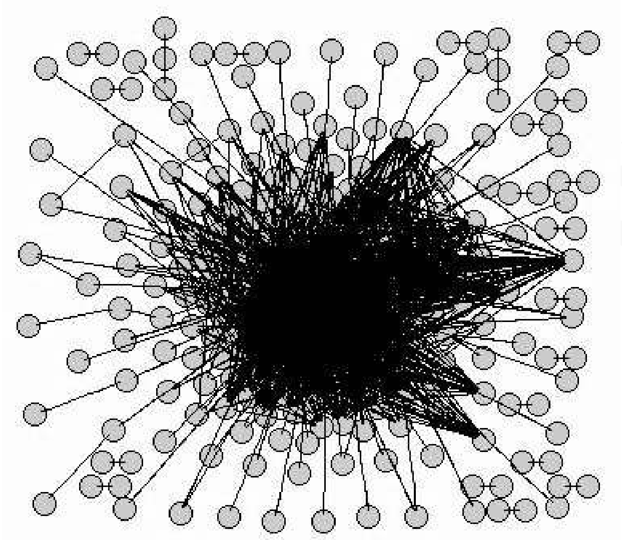


Exhibit 3

Financial Community Network Graph—February 2001—*K* = 25



first two months of 2001, all stocks were connected at the low threshold level.) The presence of one large community should not be surprising, as this is consistent with the high level of systematic risk in stock markets, as evidenced by the good fit of single-factor asset pricing models.

As threshold levels increase, we see two effects: 1) The size of the largest community drops, and 2) nodes are ejected from the largest community and become orphans, leading to more singletons. The number of singletons first increases, and then declines with threshold, as there are fewer eligible boards as well.

COMMUNITIES AND RETURNS

Differential information flow for strong and weak communities should result in differences in the mean, variances, and covariances of stock returns between the two types of communities. We compare the returns of firms within the major community to firms that do not belong to any community, i.e., singleton companies. We find that at the threshold level of $K = 5$, there are roughly the same number of singleton communities as the size of the largest community, providing a balanced comparison sample.

Comparison of Risk and Return

Using community structure to classify stocks into portfolios, we can examine the return differences between portfolios formed of community stocks and portfolios formed from singleton stocks. Exhibit 5 compares return means and standard deviations for each month in the sample period. Using all stocks within the largest community for each month, we create an equally weighted portfolio and compute its realized daily risk and return. We denote this the *community portfolio*. We also compute the risk and return for an equally weighted portfolio of singleton stocks, called the *singleton portfolio*.

The community portfolio mean return is higher than the singleton portfolio return in 14 of the 16 months in the sample. The standard deviation of returns of the community portfolio is lower than the standard deviation of the singleton portfolio in 13 of the 16 months. On average, across all months, the community portfolio provides 50 basis points per month higher return than the singleton portfolio at about half the standard deviation. Therefore, the community portfolio provides a better mean-variance trade-off.

Comparison of Covariances

The finding that the community portfolio generates better risk-adjusted return is especially striking, because stocks in the large community are likely to be more linked, given the closer information transmission among them. We want to verify whether the covariation in community stocks is higher, as would be anticipated.

To compare the covariances of the two sets of companies, i.e., the large community versus singletons, we compute the covariance matrices of returns for each set each month. We then extract the lower diagonal sub-matrix of each covariance matrix. The means of these elements are

EXHIBIT 4

Sizes of Communities Formed from Linked Message Boards

	Overlap Threshold							
Yr-MM	1	2	5	10	25	50	100	200
2000-01	25	207	409	461	392	282	148	85
	950	764	426	242	104	51	25	11
2000-02	28	212	411	474	398	286	179	93
	947	751	432	239	112	52	24	8
2000-03	16	268	566	702	567	391	228	108
	1354	1094	630	320	136	67	30	11
2000-04	17	177	425	557	448	296	153	66
	1065	897	533	254	101	48	17	8
2000-05	26	262	646	773	599	374	187	81
	1564	1322	709	322	117	52	21	5
2000-06	18	222	555	770	654	408	198	91
	1604	1394	861	419	156	66	24	8
20000-0	11	218	479	769	684	434	204	88
	1648	1437	963	454	154	67	31	14
2000-08	16	221	582	772	688	445	224	94
	1584	1379	834	404	138	63	30	11
2000-09	16	236	654	803	664	428	219	100
	1516	1294	673	319	120	56	23	8
2000-10	5	127	547	760	684	467	242	105
	1581	1459	884	466	178	75	35	17
2000-11	2	73	613	736	630	409	202	90
	1708	1637	892	473	174	75	39	20
2000-12	1	101	508	660	636	431	211	95
	1657	1555	951	549	195	75	24	13
2001-01	0	45	657	799	745	476	264	104
	1779	1734	977	534	176	90	30	14
2001-02	0	41	848	1038	734	549	259	86
	1828	1783	922	517	241	72	19	11
2001-03	4	119	569	613	506	433	220	90
	1690	1566	849	543	282	68	23	9
2001-04	8	322	668	655	534	373	176	71
	1582	1254	664	412	125	36	15	6

First number: Number of singleton communities at the given threshold level. Second number: Size of the largest of all communities.

then compared across both covariance matrices.

The results are presented in Exhibit 6. Except for one month in the dataset, there is higher covariance for community stocks.

That stocks in the large community show higher covariance supports the metric we use for connectedness in the model. Large-community stocks provide better risk-adjusted returns, despite their higher covariances.

CENTRALITY

The connectedness of the message board graph characterizes the number and the closeness of stocks in financial communities. Connectedness as based on the connections between nodes on the message board graph is a proxy for the commonality of opinion. A complementary characterization is to identify which stocks represent vortexes of information flow, captured by the notion of *centrality*.

Centrality places more emphasis on the nodes of the graph and identifies which nodes are prominent in the exchange of information, while connectedness focuses on the edges or arcs of community graphs. The more central a node on the graph, the greater its influence on other nodes.[4] From an economic point of view, stocks with higher centrality are more likely to drive the movements of other stocks, and will have greater total covariation with other stocks than with non-central stocks.

The sociological notion of centrality is developed in studies that look at power centrality in social networks (Bonacich [1972], [1987]; see also DeMarzo, Vayanos, and Zwiebel [2003] on a class of models in financial markets). We adopt this concept to quantify the centrality of information. A message board has greater information centrality if it is connected to other boards that have high centrality as well. Very high centrality may be viewed as analogous to the presence of an information hub.

The definition of centrality is in essence recursive and reflexive. Each message board's centrality is a function of every other board's centrality. Likewise, important message posters visit the important boards, which makes the boards and the posters more central.

Centrality is of interest to asset managers for two reasons: 1) By focusing on central stocks, it is possible to perceive patterns in returns that may extend to other less central stocks: and 2) it helps determine which stocks to focus on during investment analysis.

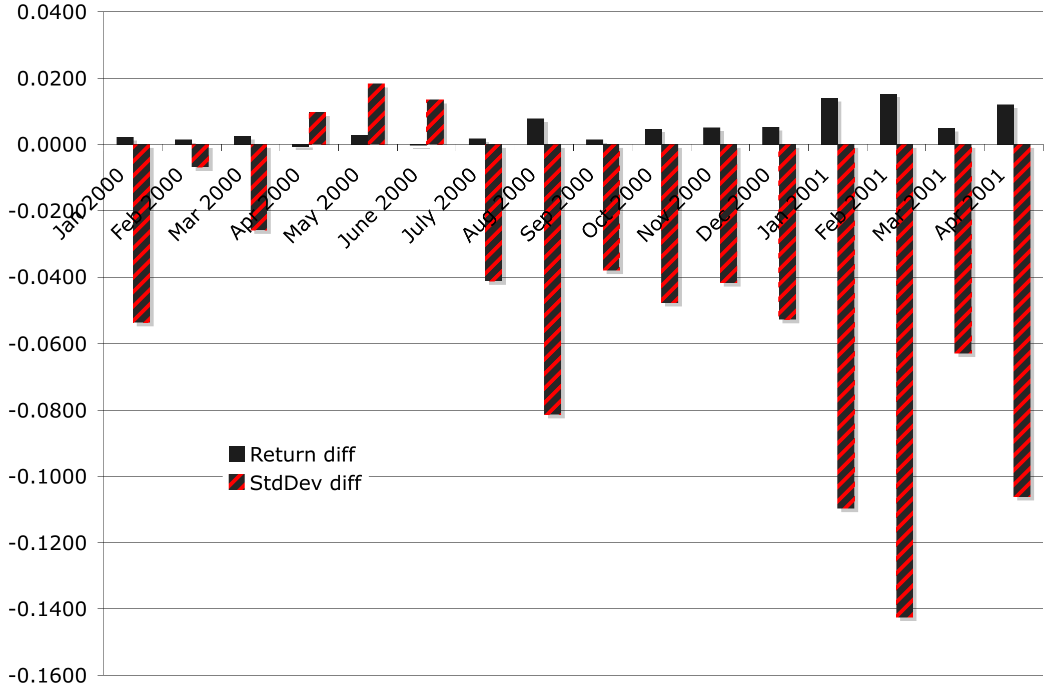
Quantifying Centrality

Given m stocks or nodes, we compute an adjacency matrix $\mathbf{A} = \{\mathbf{a}_{ij}\} \in \Re^{m \times m}$, where a_{ij} is the information overlap between stocks i and j, i.e., the number of common posters on stock message boards i, j. We define $x \in \Re^{m}$ as the vector of centrality scores. Note that centrality is a circular concept—the centrality of any one stock is a function of the centrality of all the other stocks it is connected to. Hence, we write the equation system for all centralities as follows:

$$\lambda x_i = \sum_{j \neq i} a_{ij} x_j \text{ for } \forall\, i = 1 ,\ldots, m \tag{1}$$

EXHIBIT 5

Differences in Means and Standard Deviations of Monthly Return—Large Community and Singleton Portfolio



The parameter λ is a scaling coefficient. If this is written in matrix form, we obtain

$$\lambda \mathbf{x} = \mathbf{A}\mathbf{x} \tag{2}$$

This equation parallels the definition of an eigensystem. The solution to Equation (2) provides a set of m eigenvalues λ, along with the corresponding eigenvectors x. The eigenvector corresponding to the highest absolute eigenvalue is taken to be the centrality vector of the financial community.

The centrality score for each node on the message board graph relative to any other node denotes the extent to which the message board has greater commonality of information with other message boards. Exhibit 7 depicts a few intuitive examples. An extreme example is that of a hub-and-spoke network. Assume we have a hub node and two spoke nodes. Let the weight on each spoke be 1. This would be represented by the adjacency matrix:

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{A}$$

The highest eigenvalue for this matrix is 1.4142, and the corresponding centrality score vector is [0.7071, 0.5000, 0.5000]´. As expected, hub node 1 has a higher score than the spoke nodes 2 and 3.

On the other hand, a triangular network, depicted in the matrix:

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \tag{B}$$

yields a centrality vector of [0.5774, 0.5774, 0.5774]´, i.e., all nodes have the same centrality.

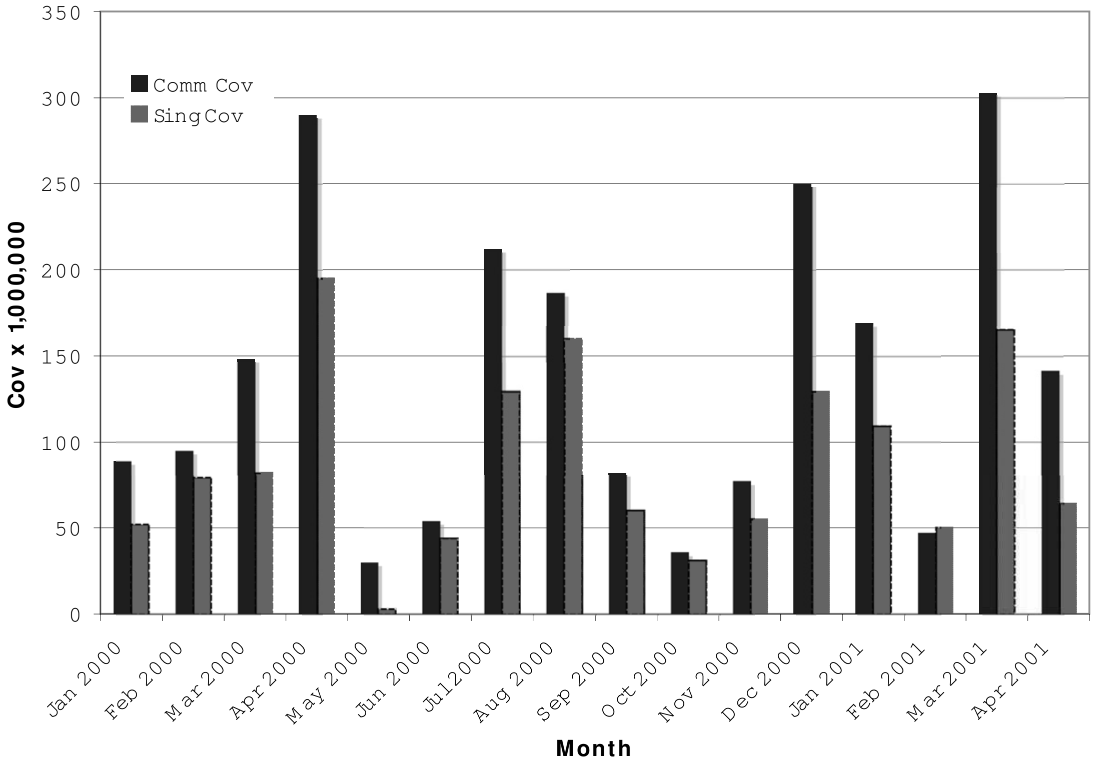
As a final example, consider an unbalanced triangular network with the adjacency matrix:

$$\begin{bmatrix} 0 & 2 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{C}$$

The centrality vector is [0.7071, 0.6325, 03162]´, which shows that node 3 is less connected than nodes 1 and 2.

Exhibit 7 graphs the systems represented in these matrices. The three diagrams relate to the matrices (A)-(C).

EXHIBIT 6
Average Covariance in Month



For each month in the data set, we compute the centrality scores for all message boards, and find that a few boards appear to have very high centrality. Exhibit 8 provides a visual depiction of centrality and connectedness of the data for January 2000. The location of a ticker on the graph is a function of the degree of connectedness. The most highly connected stocks are in the center of the graph, and the distance of a stock from the center reflects declining connectedness. The distances are scaled to reflect the standard deviation of the sample on a (–1, +1) grid.

Centrality is reflected by the size of the ticker symbol on the graph; the top third of the tickers are large; the middle group are medium-sized; and the bottom third are small. To avoid clutter on the graph, we reflect connectedness for a community threshold level of $K = 50$. This makes the number of tickers on the graph more visible.

The plot shows that the two most connected stocks are Lucent (LU) and America On-Line (AOL). Other stocks in the close vicinity are Compaq (CPQ), AT&T (T), and International Business Machines (IBM). These stocks have high centrality too. There are other stocks with high centrality that do not have high connectedness, such as SBC Communications (SBC), Citicorp (C), and Motorola (MOT). Stocks with low centrality do not appear to have high connectedness.

There do not appear to be strong industry-based concentrations, except that the topmost centrality firms are technology companies. Even at the threshold level of $K =$ 50, there are many non-high-tech firms such as Bristol Myers (BMY), Toys-R-Us (TOY), Coca-Cola (KO), Boeing (BA), Abbott Labs (ABT), and McDonald's (MCD), which reflects the fact that the rankings of connected firms and central firms may be driven more by broad-based small investor interest than by industry concentration.

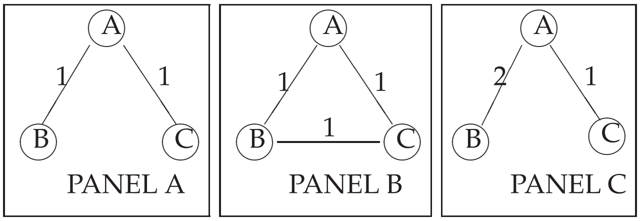
Centrality and Return Covariance

The more central a stock is, the more closely it is tied by information links to other stocks. We can evaluate economic meaningfulness by examining the return covariance of hub stocks with others. Whether portfolio managers should focus on more central stocks is related to whether centrality is indeed reflected in a higher covariance of central stocks with other stocks.

We examine this proposition as follows. First, we

EXHIBIT 7

Three Matrices in Centrality Section



compute the centrality measure x_i for all firms $i = 1, \ldots, m$, as in Equation (1). Next, we compute the covariance matrix $S = \{s_{ij}\}$ of returns for the m firms each month. Then, for each firm i, we compute the average pairwise covariance with all the other firms:

$$c_i = \frac{1}{m-1} \sum_{j \neq i} s_{ij} \text{ for } \forall i = 1\,, \ldots, m$$

Finally, we compute the correlation between centrality and return covariance:

$$Corr[x_i, c_i] \text{ for } \forall i = 1, \ldots, m$$

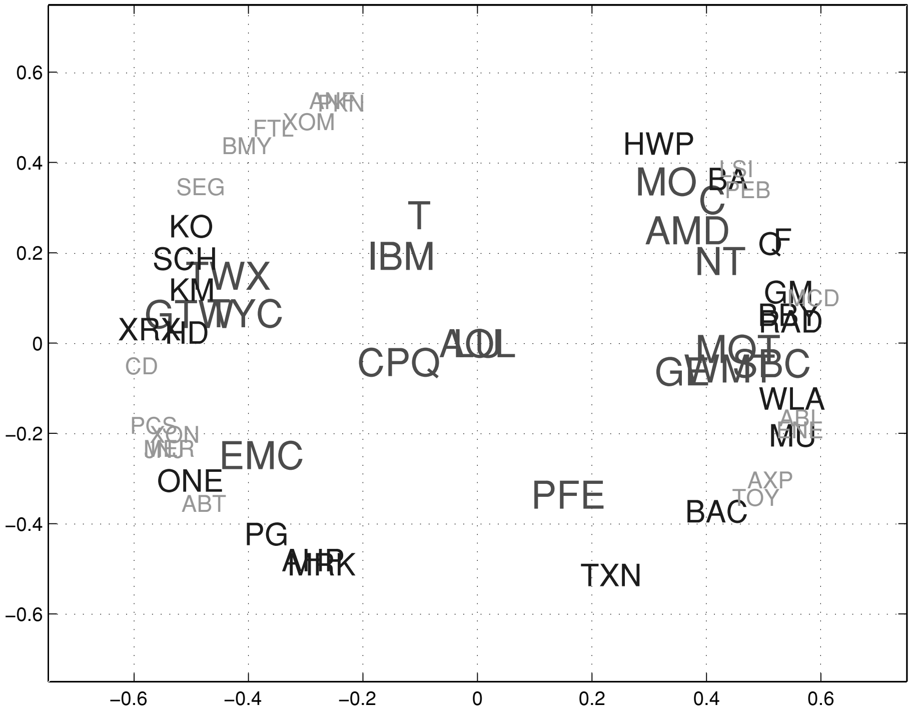
If this correlation is positive, it suggests that a focus on high-centrality stocks may be justified in portfolio analysis.

The results are portrayed in Exhibit 9. We plot the correlation value for each month in our sample for both the community portfolio and the singleton portfolio. Exhibit 9 shows that most of the outcomes are positive, which supports the idea that stocks are positively correlated with the returns of central stocks.

These results on the relationship between returns and connectedness and centrality suggest that Internet discussion of stocks is related to the way information is impounded into stock prices. Antweiler and Frank [2002] find that the extent of discussion on individual stock message boards is related to volatility and returns. Our findings complement this work at an aggregate level. We find there are economically meaningful reasons to study the sociology of investors.

EXHIBIT 8

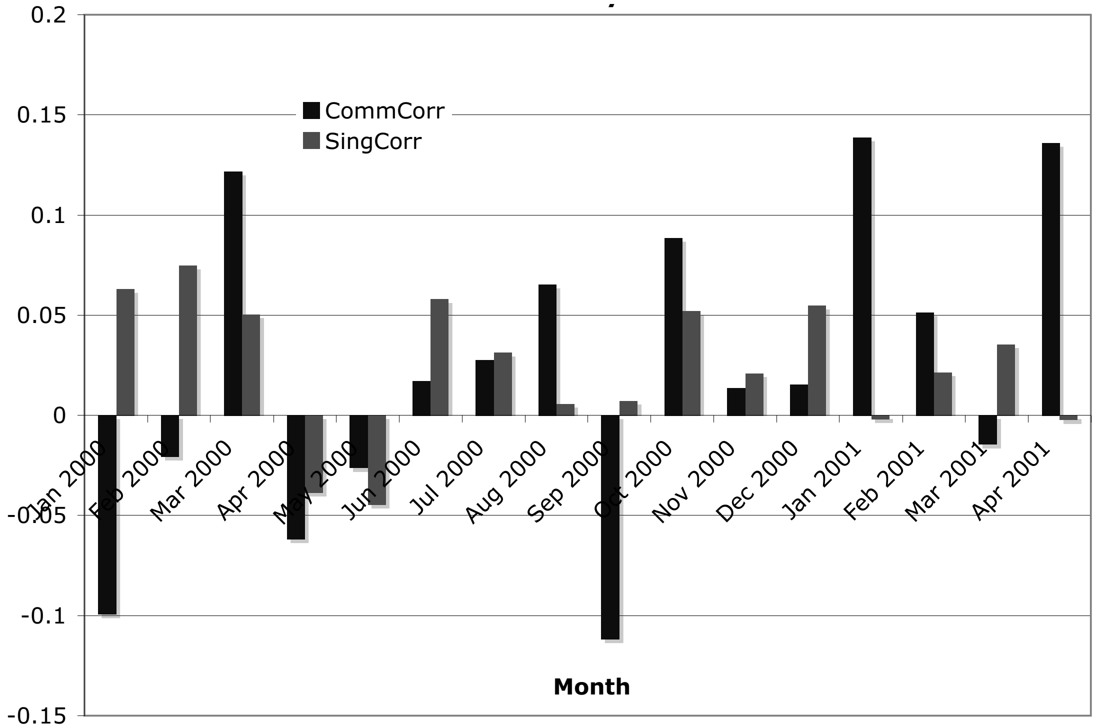
Visual Depiction of Connectedness and Centrality—January 2000—*K* = 50



Visual Depiction of Connectedness and Centrality: This plot presents a vi-

EXHIBIT 9

Correlation of Centrality with Return Covariance



SUMMARY

Many theories in finance assume a strong link between stock returns and information, yet its social mechanics have been relatively unexplored. We use millions of messages posted to stock market discussion forums to understand how opinions are linked across tickers during small investor discussion. We thus define a new information unit, the *financial community*, clusters of tickers sharing and accessing the same information generators.

Baker [1984] developed a sociological study of markets two decades ago by examining the network of traders on the floor of the options exchange. He found that the forces of bounded rationality and opportunism led to sociological distortions from rational markets. For example, as the connectedness of the network increased, stock volatility actually increased, contrary to what one would expect in a hyperrational market.

We do not find this effect in our study. In our setting, as the connectedness in the financial community increases, there is less volatility in the highly connected portfolio (despite an increase in covariances of the components). Many other features that we find are consistent with those Baker describes. For instance, he hypothesizes that an increase in network size results in greater similarity of behavior across the network, which takes the form of higher means and lower standard deviations among community stocks.

Our graph-theoretic techniques used to detect financial communities and to summarize their properties give portfolio managers an alternative portfolio classification scheme. We find that strong community stocks display *connectedness*: higher mean returns and lower standard deviations of returns than unconnected stocks in weak communities. The superior risk-return trade-off in community stocks suggests a portfolio strategy that is long community stocks and short singleton stocks would be fruitful. The greater the extent of connectedness in a financial community, the greater the covariation of returns within the community as opposed to covariations among stocks that are not part of a major financial community. This suggests that more diversification comes from singleton stocks.

Finally, following eigenvector techniques, we detect stocks that are hubs for information flow, using a *centrality* measure. We find that stocks with high centrality scores tend to have greater covariation with other stocks than those with low scores. This suggests analysts might gainfully focus on high-centrality stocks.

We conclude that network analysis of financial communities provides a way to base portfolio strategies on information flows.

APPENDIX

Formal Graph Definition

Our community graph is denoted $G = [V, E]$, where V is a node set, and E is an edge (connections) set. We have exactly one node per stock ticker, and the number of nodes is m, i.e., $|V| m$. Canonical notation for the edges will be $e = (v, w) \in E$, where $v, w \in V$ are nodes, and the size of the edge set is $|E| = n$. It will be clear from the context whether the edge set implies a directed or undirected graph. Graph G depicts the strengths of connections between tickers.

Our graph is *message handle-based*. The handle $h \in H$ of a message is simply the poster's screen name. H is the node set of handles. Edge weights between nodes are a count of the number of common handles between two message boards.

We define a metric $p(e)$ for each edge, such that

$$p(e) = p(v, w) = p(w, v) = \text{Count}_H\{(h \in v) \cap (h \in w)\}$$

Therefore, $p(e)$ is the count of common posters on two boards. Under this metric, the graph is undirected.[5]

The graph may be expressed in the form of an $(m \times m)$ *adjacency* matrix, such that cell values in the matrix $A(v, w) = p(v, w)$. We focus on a graph based on message handles, i.e., distinct poster screen names. Since the graph is undirected, the adjacency matrix $A(v, w)$ is square-symmetric.

ENDNOTES

The authors are grateful to David Leinweber for both the data and his comments; and Robert Hendershott, Ravi Jagannathan, and Meir Statman for their comments. Thanks also to Codexa, Inc., for the data. Das gratefully acknowledges support from a Breetwor Fellowship and the Dean Witter Foundation. Sisk was a graduate student at UCLA when this project was commenced, and a researcher at Leinweber & Co when most of the work was undertaken.

[1] One could imagine other ways connection strengths are determined, such as the number of times IBM is mentioned on Google's message board. In this case, the adjacency matrix would no longer be symmetric.

[2] We can count the number of communities in our universe of message boards very quickly by running a depth-first search (DFS) through the adjacency matrix A. The depth-first search begins from any node and circumscribes a community by working through and accumulating under one community all connected nodes on the graph until it encounters no further connected nodes to visit. These are standard and basic algorithms; see Tarjan [1983]. DFS then finds the next community by restarting the search from any unvisited node, visiting as many connected nodes as possible. The number of distinct communities is equal to the number of times the DFS restarts from an unvisited node. The algorithm is fast and takes $O(m + n)$ work only, (where m is the number of nodes, and n is the number of edges), i.e., it is a linear complexity algorithm. During the DFS run, we simultaneously determine the node clusters that constitute the communities or connected components of the graph. This does not change the run time of the algorithm.

[3] This finding is consistent with the metaphor of six degrees of separation, i.e., the notion that everything is more connected than we expect.

[4] See Theorem 1 in DeMarzo, Vayanos, and Zwiebel [2003] for a formal presentation of this idea.

[5] Another example of a graph scheme is ticker reference-based. In this scheme, a metric for edge weights $t(v, w)$ is defined as the extent to which a message board for ticker v has mention of ticker w. Thus:

$$t(e) = t(v, w) = \text{Count}_v\{\text{No. of Msgs with References to node } u$$

Note that this graph is directed. Therefore, $t(v, w) \neq t(w, v)$. The count is taken over the number of messages that contain cross-references. Multiple cross-references within the same message are not double-counted. We can also compute another version of this metric for undirected graphs, where the edge weights are $t(v, w) + t(w, v)$.

REFERENCES

Admati, Anat, and Paul Pfleiderer. "Noisytalk.com: Broadcasting Opinions in a Noisy Environment." Working paper, Stanford University, 2001.

Antweiler, W., and Murray Frank. "Internet Stock Message Boards and Stock Returns." Working paper, University of British Columbia, 2002.

——. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance*, v59(3) (2004), pp. 1259-1295.

Bagnoli, Mark, Messod Beneish, and Susan G. Watts. "Whisper Forecasts of Quarterly Earnings per Share." *Journal of Accounting and Economics*, v28(1) (1999).

Baker, Wayne E. "The Social Structure of a National Securities Market." *American Journal of Sociology*, v89(4) (1984), pp. 775-811.

Barber, Brad M., and Terrance Odean. "All that Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors." Working paper, University of California at Davis, 2002.

Barberis, Nicholas, and Andrei Shleifer. "Style Investing." *Journal of Financial Economics*, v68(2) (2003), pp. 161-200.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. "Informational Cascades and Rational Herding: An Annotated Bibliography." Working paper, UCLA/Anderson, The Ohio State University, and Yale School of Management, 1996.

Bonacich, Phillip. "Power and Centrality: A Family of Measures." *American Journal of Sociology*, v92(5) (1987), pp. 1170-1182.

——. "Technique for Analyzing Overlapping Memberships." *Sociological Methodology*, v4 (1972), pp. 176-185.

Boudoukh, Jacob, Matthew Richardson, and Robert Whitelaw. "A Tale of Three Schools: Insights on Autocorrelations of Short-Horizon Stock Returns." *The Review of Financial Studies*, v7(3) (1994), pp. 539-573.

Cornell, Bradford. "Comovement as an Investment Tool." *The Journal of Portfolio Management*, Spring 2004, pp. 106-111.

Das, Sanjiv, and Mike Y. Chen. "Yahoo! For Amazon: Opinion Extraction from Small Talk on the Web." Working paper, Santa Clara University, 2000.

Das, Sanjiv, Asis Martinez-Jerez, and Peter Tufano. "e-Information." Working paper, Harvard Business School, 2001.

DeMarzo, Peter, Dimitri Vayanos, and Jeffrey Zwiebel. "Persuasion Bias, Social Influence, and Uni-Dimensional Opinions." *Quarterly Journal of Economics*, v118 (2003), pp. 909-968.

Harris, M., and A. Raviv. "Differences of Opinion Make a Horse Race." *The Review of Financial Studies*, v6 (1993), pp. 473-506.

Kyle, A. "Continuous Auctions and Insider Trading." *Econometrica*, v53(6) (1985), pp. 1315-1335.

Tarjan, Robert E. "Data Structures and Network Algorithms." *CBMS-NSF Regional Conference Series in Applied Mathematics*, 1983.

Tumarkin, Robert, and Robert Whitelaw. "News or Noise? Internet Message Board Activity and Stock Prices." *Financial Analysts Journal*, v57 (2001), pp. 41-51.

Watts, Duncan. "A Simple Model of Global Cascades on Random Networks." *Proceedings of the National Academy of Sciences*, 99-9, April 2002, pp. 5766-5771.

Welch, Ivo. "Sequential Sales, Learning, and Cascades." *Journal of Finance*, v47(2) (1992), pp. 695-732.

Wysocki, Peter. "Cheap Talk on the Web: The Determinants of Postings on Internet Stock Message Boards." Working paper, 1999.

Zuckerman, Ezra, and Hayagreeva Rao. "Shrewd, Crude or Simply Deluded? Comovement and the Internet Stock Phenomenon." Working paper, MIT, 2003.

To order reprints of this article, please contact Ajani Malik at amalik@iijournals.com or 212-224-3205.