Three Essays on Applying Structure Analysis in Financial Econometrics

by

Keyi Zhang

M.A., Simon Fraser University, 2011
B.Econ., Beijing Normal University, 2007

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Economics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Fall 2018

Approval

Name: **Keyi Zhang**

Degree: **Doctor of Philosophy (Economics)**

Title: **Three Essays on Applying Structure Analysis in Financial Econometrics**

Examining Committee:

Chair: Brian Krauth
Associate Professor

Ramazan Gençay
Senior Supervisor
Professor

Bertille Antoine
Supervisor
Associate Professor

Andrey Pavlov
Internal Examiner
Professor
Beedie School of Business

Nikola Gradojevic
External Examiner
Professor
Department of Economics and Finance
University of Guelph

Date Defended: **November 23, 2018**

Abstract

This thesis is composed of three essays on applying different structure analyses in financial econometrics. The first chapter, entitled “Application of Wavelet-based Structures in Time-Series Index Forecasting”, is based on a joined work with Ramazan Gençay and M. Ege Yazgan, which is published in *Economics Letters* in 2017. This essay explores the potential of wavelet-based multiresolution analysis in forecasting. A hierarchical structure for a single time series index is defined and estimated in frequency domain, based on which a forecast combination technique is applied to achieve an improvement in forecast accuracy.

The second chapter, entitled “Application of Network Structures in Stock Return Volatility Forecasting”, is based on a joined work with Xiao Yu and Ramazan Gençay. This essay explores the potential of network analysis in forecasting stock return volatility. A customer and supplier network structure is identified and incorporated in the usual reduced form stock return volatility model. Results show that there is a propagation dynamic of stock return volatility along supply chain, and incorporating customer channel improves the accuracy of volatility forecasting.

The third chapter, entitled “Application of Network Structures in Mutual Fund Performance Forecasting”, is based on a joined work with Ramazan Gençay, which is published in Singapore Economic Review in 2018. This essay explores the relationship between mutual fund performance persistence and the network structure of mutual funds. By constructing a network of mutual funds based on the commonality of their stock holdings, we can identify mutual funds that are more likely to possess momentum in performance.

Keywords: Network analysis; Wavelet decomposition; Forecasting; Financial market

Table of Contents

List of Tables

List of Figures

Chapter 1

Application of Wavelet-based Structures in Time-Series Index Forecasting

Wavelet-based multiresolution analyses can decompose a time series into a set of constitutive series with an explicitly defined hierarchical structure. In this paper, we show that this decomposition method can improve the accuracy of forecasts of original times series data.

A hierarchical time series includes multiple times series in which the high-level observations are aggregated according to low-level data. Economic data often have this hierarchical structure. For example, GDP data for a country, state, and city are a group of hierarchical time series based on geography. Conventional approaches to performing forecasts using such hierarchical data involve either a top-down or bottom-up method or a combination of the two. The top-level data could be forecast first, and then these forecasts could be disaggregated based on historical proportions (top-down approach); alternatively, the bottom-level data could be forecast first, and then additional data could be included to obtain the top-level forecasts (bottom-up approach).[1] Thus, when performing forecasts, the value of the data and the structure is important.

Ignoring the hierarchical structure of the data and forecasting all series at all levels independently will usually lead to the undesirable consequence in which higher-level forecasts are not equal to the sum of the directly related lower-level forecasts. To address this issue, Hyndman, Ahmed, Athanasopoulos, and Shang (2011) presents a framework to ensure that forecasts are added appropriately by adjusting the independent forecasts. That is, given multiple times series that are hierarchically organized, an unbiased and efficient forecast can be achieved without losing the hierarchical structure.

In this paper, we extend the application of Hyndman et al. (2011) to any univariate times series data. We apply wavelet-based multiresolution analyses to expand univariate time series data into a group of hierarchical series in a meaningful manner. This application provides the opportunity to study and apply the structure of the data when forecasting.

To examine whether the wavelet decomposition can improve forecasting accuracy, we compare the forecast accuracy obtained by different methods. The accuracy benchmarks are forecasts performed using conventional univariate models, which are applied to the wavelet-decomposed hierarchical series, and the forecasts at the top level are reconciled as the raw data. Therefore, for each univariate model, a pair of forecast results is obtained, with one based solely on raw data and one based on a multiresolution analysis. A comparison of these results can be used to test the accuracy of our method because the forecast model is not changed and any improvements in forecast accuracy can be attributed to the wavelet decomposition method.

In multiresolution analyses, sequences of local averages ("smooths") and differences ("details") in the time series at difference scales are obtained. Each smooth and detail has the same number of observations as the raw time series data, and an explicit hierarchical structure is observed among the raw data and the decomposed smooths and details.[2] We apply the traditional univariate time series prediction methods to the raw data and their decomposed components independently, then combine and reconcile these forecasts according to the hierarchical structure so that the revised forecasts will preserve the same hierarchy as the original sample.[3] The proposed methodology is parsimonious because the forecast is based only on the time series of interest, and it is independent of the forecasting models. We demonstrate our proposed method by forecasting the daily returns of the S&P500 index using a sample from February 14, 2011 to August 19, 2016. Several benchmark univariate time series prediction methods are considered. Compared with the approach that applies only conventional methods to raw data, the forecast accuracy is improved when the wavelet-based multiresolution analysis is applied. These results suggest that wavelet decomposition can help improve the accuracy of forecasting a univariate time series.

1.1 Method

Our purposed methodology is based on the approach detailed in Hyndman et al. (2011) and augmented by a wavelet-based multiresolution analysis. Essentially, we use a regression model (OLS) to reconcile the forecasts at different wavelet-decomposed series, which maintains consistency among the refined predictions at all levels. Therefore, higher-level predictions are equal to the sum of their directly related lower-level predictions.

The general procedure takes four steps. First, we apply wavelet-based multiresolution analysis to decompose the raw data X into wavelet details and smooths, $\{S_1, D_1, S_2, D_2\}$. These multiple time series have the properties of a group of hierarchical time series.

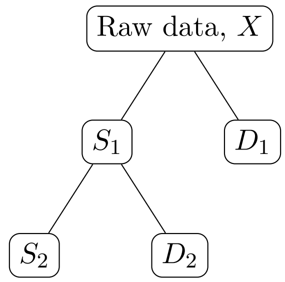


Figure 1.1: **Example of a Hierarchical Structure**

Next, for a given training sample with size T, we can use a conventional univariate model to conduct first-round forecasts at all hierarchical levels independently with a horizon h. For example, $\hat{Y}_{T+h} = \left[\hat{X}, \hat{S}_1, \hat{D}_1, \hat{S}_2, \hat{D}_2\right]'$ can be predicted by $AR(1)$, where $\hat{Y}_{T+h}$ has a dimension 5-by-1 and time index $T+h$.

Because the wavelet-decomposed components (raw data, details and smooths at different levels) have an explicit hierarchical structure, we can express the variables on each scale as a linear combination of the lowest level (base-level) variables, $\beta = [S_2, D_2, D_1]'$, which have no descendants. For example, $X = S_2 + D_2 + D_1$, and $S_1 = S_2 + D_2$. Therefore, a "summing" matrix, Z, with entries of $\{0, 1\}$ can fully capture this linear relationship in a given hierarchy.

$$Y \equiv \begin{bmatrix} X \\ S_1 \\ D_1 \\ S_2 \\ D_2 \end{bmatrix} = \begin{bmatrix} S_2 + D_2 + D_1 \\ S_2 + D_2 \\ D_1 \\ S_2 \\ D_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} S_2 \\ D_2 \\ D_1 \end{bmatrix} \equiv Z\beta$$

With most univariate time series models[4], the first-round independent forecasts, $\hat{Y}_{T+h} = \left[\hat{X}, \hat{S}_1, \hat{D}_1, \hat{S}_2, \hat{D}_2\right]'$, do not preserve the same hierarchical structure as the original training sample, Y, i.e., $\hat{Y}_{T+h} \neq Z\hat{\beta}_{T+h}$. Thus, those forecasts are not inherently consistent with each other.

Because of the independent predictions at different hierarchical levels and the multiresolution structure, we regress the prediction result vector, $\hat{Y}_{T+h}$, on the corresponding summing matrix, Z. The objective of this step is to identify the set of base-level forecasts, $\tilde{\beta}_{T+h} = \left[\tilde{S}_2, \tilde{D}_2, \tilde{D}_1\right]'$, that minimizes the squared deviation from the first round of independent forecasts.

$$\tilde{\beta}_{T+h} = \underset{b}{\operatorname{argmin}} \left(\hat{Y}_{T+h} - Zb\right)' \left(\hat{Y}_{T+h} - Zb\right)$$

Thus, the refined base-level forecasts, $\tilde{\beta}_{T+h}$, can be estimated using OLS.[5]

Finally, the refined optimal combination forecasts[6] at all hierarchical levels can be calculated by $\tilde{Y}_{T+h} = Z\tilde{\beta}_{T+h}$. Among these forecasts, we are particularly interested in the accuracy of the refined forecasts at the top level, $\tilde{X}_{T+h} = \tilde{S}_{2,T+h} + \tilde{D}_{2,T+h} + \tilde{D}_{1,T+h}$, which corresponds to the original time series, X.

1.2 Empirical

1.2.1 Data

A sample of S&P 500 index daily observations from February 14, 2011 to August 19, 2016 are used to demonstrate our forecast results. The daily returns are calculated according to the first difference in the log-transformed index price level on each trading day. The prediction is conducted using a multi-step horizon out-of-sample forecast approach without model re-estimation. No information beyond each training sample is needed to generate the corresponding forecasts. We apply a rolling window approach to generate a sequence of similar out-of-sample predictions.

1.2.2 Results

Because the goal of this paper is to demonstrate the marginal contribution obtained by applying the wavelet decomposition approach, the accuracy of the forecasts should be compared between the methods with and without the wavelet-based multiresolution analysis using the same univariate time series model. Therefore, any forecast improvement can be attributed to the wavelet decomposition instead of to a specific time series model. In addition, cross-model comparisons are not directly relevant to our purpose, which explains why the forecast results between $AR(1)$ and $AR(2)$, which are both augmented by a multiresolution analysis, are not compared.

We demonstrate the forecast improvements using the mean absolute error (MAE) and the root mean squared error (RMSE). The benchmark univariate time series models include an autoregressive model with different numbers of the lag terms, $ARMA(1,1)$ and an optimal ARIMA model that uses the automatic algorithm of Hyndman and Khandakar (2008). Different lengths of rolling windows and forecast horizons are examined. When calculating the multi-step forecasts, $\hat{y}_{T+h}$ with $h > 1$, the predicted values from shorter steps are used, $\hat{y}_{T+\tau}$ with $\tau < h$.

In tables 1.1 and 1.2, we show the percentage improvements (decreases) in the MAE and RMSE of the multi-step forecasts performed using the multiresolution-analysis augmented method and compare these results with the forecasts obtained using the corresponding conventional approaches. Multiresolution analyses are conducted using a level-2 MODWT with a haar wavelet and a periodic boundary condition. The benchmark models include $AR(1)$, $AR(2)$, $AR(3)$, $AR(6)$, $ARMA(1,1)$, and $ARMA.auto$, which is based on an optimal ARIMA model using the automatic algorithm to select the model specifications (Hyndman and Khandakar, 2008). We consider different forecast windows (10 days, 30 days, and 50 days)[7] and forecast horizons (maximum of 14 days).

In most cases, significant forecast improvements are obtained, and an interesting pattern is observed. The improvement in forecast accuracy obtained from using the proposed method is more significant in cases with a shorter estimation window or longer forecast horizon. For example, in the $AR(2)$ model, a 45.71% decrease in the MAE is observed when a 10-day window and a 14-day forecast horizon are used, whereas a 7.44% decrease in the MAE is observed when a one-day forecast horizon is used and a 0.30% decrease is observed when a 50-day estimation window is used. This pattern suggests that the proposed method is relatively more powerful when the forecast task is more demanding (fewer observations or a longer forecast horizon). The statistical significance of the forecast improvements are shown in the Table 1.3 and 1.4.

As a robustness check, we also conducted the empirical test using S&P 500 Net Total Return index, which reflects the effects of the dividend reinvestment and withholding tax deduction. Similar results are shown in Table 1.5 - 1.8.

In figures 1.2 and 1.3, we show the side-by-side comparisons of the MAE and RMSE between the results of the multiresolution-analysis augmented method and the corresponding conventional approaches: $AR(1)$, $AR(2)$, and $ARMA(1,1)$. For all cases, the forecast accuracy deteriorates as the forecast horizon is increased, whereas our proposed method shows consistently better results (lower prediction error in terms of the MAE and RMSE) than the benchmark. This result suggests that the multiresolution-analysis augmented method is a relatively more robust method for increasing forecast horizons.

1.3 Conclusions

Hyndman et al. (2011) shows that optimal combination forecast results perform well compared with the results of conventional top-down approaches and the bottom-up method. However, these optimal combination forecasts rely on the availability of multivariate time series that have a hierarchical structure and can be aggregated at different levels.

In this study, we propose a methodology for extending the forecast approach in Hyndman et al. (2011); Hyndman, Lee, and Wang (2016) to any univariate time series data. We use wavelet-based multiresolution analyses to decompose the original univariate time series into a group of time series (wavelet details and smooths) that have an explicit hierarchical structure. Therefore, the optimal combination forecast approach can be applied to a univariate times series. We demonstrate the improved accuracy of the multiresolution-analysis augmented forecasting method using an example from the S&P500 index.

1.4 Notes

[1]The examples of combination forecasts include Claeskens, Magnus, Vasnev, and Wang (2016); Del Negro, Hasegawa, and Schorfheide (2016); Chen, Turnovsky, and Zivot (2014); Zellner and Tobias (2000); Rapach, Strauss, and Zhou (2010); Fliedner (1999); Kohn (1982); Tiao and Guttman (1980).

[2]More technical details on wavelet-based multiresolution analyses can be found in Gençay, Selçuk, and Whitcher (2001).

[3]Our approach using wavelet transform for forecasting is similar to Conejo, Plazas, Espinola, and Molina (2005); however, the main difference is that in Conejo et al. (2005), the forecast is conducted via a bottom-up approach, whereas we apply an optimal reconciliation approach that utilizes information from all levels.

[4]One exception is the random walk model in which prediction values are simply the last observations of the training sample.

[5]Unequal weights can also be used in estimating betas. The performance of alternative estimation methods (such as WLS or FGLS) shall be evaluated in future studies.

[6]The term "optimal combination" is from Hyndman et al. (2011). The reconciled forecast is considered as optimal in the sense that it minimizes the mean squared deviation from the independent predictions at all hierarchy levels.

[7]$AR(6)$ is not used in the cases with 10-day forecast windows due to the limit of the sample size.

1.5 Tables and Figures

Table 1.1: **Forecast Improvement (MAE) for S&P500 Index Return with Different Horizons**

The numbers in the table are percentage improvements (decreases) in mean absolute error (MAE) of multi-step forecasts by multiresolution-analysis augmented method, comparing with forecasts using corresponding benchmark models without using multiresolution analysis. Multiresolution analysis is conducted by level-2 MODWT with haar wavelet. The sample is based on the daily returns of the S&P500 index from February 14, 2011 to August 19, 2016. Rolling window forecasts are conducted with 10, 30, and 50 days as the window length. Each forecast practice is based on the information limited to the observations during corresponding window period. For example, the procedure for an h-step 10-day rolling window forecast includes: 10-day observations are used to generate wavelet-based hierarchical multiple time series; an h-step out-of-sample forecast is conducted for each times series without re-estimation, where multi-step forecasts, $\hat{y}_{T+h}$ with $h > 1$, take previous period's predicted values as input; refined combination forecasts are calculated based on $\hat{y}_{T+h}$ at all levels. The last column are the forecast improvements based on the average MAE among all h-step forecasts. The $ARMA.auto$ is based on an optimal ARIMA model using the automatic algorithm of Hyndman and Khandakar (2008).

	Forecast horizon, daily basis														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave
10-day rolling window forecast															
$AR(1)$-MRA	2.02	−1.43	0.40	−0.15	0.77	2.20	4.01	6.12	10.10	13.96	19.50	24.73	30.11	34.22	16.85
$AR(2)$-MRA	7.44	2.19	6.32	6.13	14.20	16.80	27.78	34.57	39.78	42.52	44.28	44.99	45.44	45.71	45.16
$AR(3)$-MRA	13.66	7.01	13.28	12.14	16.14	17.64	19.50	17.53	16.20	9.94	3.94	−6.51	−20.57	−39.17	−29.4
$ARMA(1,1)$-MRA	5.54	2.75	4.27	3.51	5.34	5.57	8.04	10.43	12.42	15.52	18.44	22.44	26.12	28.72	15.34
$ARMA.auto$-MRA	−1.57	−1.24	−0.46	−1.83	−2.09	−1.13	−1.46	−0.90	−2.01	−1.69	−1.73	−1.49	−2.89	−2.86	−1.67
30-day rolling window forecast															
$AR(1)$-MRA	−0.26	−1.62	−1.08	−1.13	−0.40	−0.74	−0.13	−0.80	−0.26	−0.14	−0.53	−0.94	−0.53	−0.34	−0.64
$AR(2)$-MRA	0.42	−1.02	0.04	0.10	0.93	0.65	1.34	1.14	1.36	2.13	3.06	2.33	2.70	3.64	1.36
$AR(3)$-MRA	1.55	0.52	1.29	0.59	1.87	0.54	0.79	0.61	1.32	1.57	2.75	3.20	2.93	3.43	1.67
$AR(6)$-MRA	5.77	2.35	2.25	0.85	2.36	2.98	3.75	2.67	4.72	5.11	8.23	8.88	7.90	9.36	4.97
$ARMA(1,1)$-MRA	0.91	0.53	0.55	0.41	2.02	2.27	2.69	2.47	2.70	2.80	4.15	4.18	4.72	4.52	2.52
$ARMA.auto$-MRA	−2.88	−0.44	−0.85	−1.00	−0.67	−0.54	0.36	0.24	0.33	0.45	0.64	0.98	0.15	0.50	−0.20
50-day rolling window forecast															
$AR(1)$-MRA	−1.63	−1.70	−1.03	−1.12	−0.41	−0.47	−0.48	−0.44	−0.16	−0.14	−0.31	−0.39	−0.33	−0.17	−0.63
$AR(2)$-MRA	−1.95	−0.58	−0.33	−0.24	−0.45	0.00	0.31	0.22	0.25	0.15	0.39	−0.06	0.16	0.30	−0.13
$AR(3)$-MRA	−2.59	−0.31	0.71	−0.83	−0.45	−0.22	0.33	0.26	−0.36	0.06	0.58	0.22	0.06	0.38	−0.16
$AR(6)$-MRA	−0.87	0.07	0.42	−0.70	−0.06	0.62	0.74	0.71	1.81	1.59	2.35	1.74	2.29	2.76	0.95
$ARMA(1,1)$-MRA	−3.08	0.41	−0.53	1.25	0.85	0.32	0.65	1.18	1.89	1.23	1.04	1.64	1.38	2.10	0.74
$ARMA.auto$-MRA	−4.83	−2.42	−1.57	−1.81	−2.13	−0.96	−0.36	−0.43	0.60	−0.55	0.10	−0.59	−0.63	−0.80	−1.18

Table 1.2: **Forecast Improvement (rMSE) for S&P500 Index Return with Different Horizons**

The numbers in the table are percentage improvements (decreases) in root mean squared error (rMSE) of multi-step forecasts by multiresolution-analysis augmented method, comparing with forecasts using corresponding benchmark models without using multiresolution analysis. Multiresolution analysis is conducted by level-2 MODWT with haar wavelet. The sample is based on the daily returns of the S&P500 index from February 14, 2011 to August 19, 2016. Rolling window forecasts are conducted with 10, 30, and 50 days as the window length. Each forecast practice is based on the information limited to the observations during corresponding window period. For example, the procedure for an h-step 10-day rolling window forecast includes: 10-day observations are used to generate wavelet-based hierarchical multiple time series; an h-step out-of-sample forecast is conducted for each times series without re-estimation, where multi-step forecasts, $\hat{y}_{T+h}$ with $h > 1$, take previous period's predicted values as input; refined combination forecasts are calculated based on $\hat{y}_{T+h}$ at all levels. The last column are the forecast improvements based on the average rMSE among all h-step forecasts. The $ARMA.auto$ is based on an optimal ARIMA model using the automatic algorithm of Hyndman and Khandakar (2008).

	Forecast horizon, daily basis														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave
10-day rolling window forecast															
$AR(1)$-MRA	3.62	0.25	4.06	8.37	17.53	26.67	34.68	38.75	41.52	42.71	43.63	44.17	44.62	44.92	42.90
$AR(2)$-MRA	11.02	9.98	28.37	37.57	45.56	45.30	46.28	46.02	46.19	46.13	46.16	46.15	46.15	46.15	46.15
$AR(3)$-MRA	17.64	17.78	36.08	37.89	43.14	41.38	41.24	40.01	38.31	35.08	29.69	20.82	6.90	−13.86	−4.00
$ARMA(1,1)$-MRA	6.14	3.55	8.93	9.40	17.11	24.44	31.37	37.02	40.84	43.92	44.41	45.38	45.78	45.96	41.51
$ARMA.auto$-MRA	0.30	0.48	0.53	−0.59	−1.46	0.84	0.68	2.10	0.20	2.89	−0.12	1.70	1.75	0.51	0.71
30-day rolling window forecast															
$AR(1)$-MRA	0.70	−1.11	−0.55	−1.12	0.14	−0.52	0.43	−0.40	−0.04	0.08	−0.98	−1.65	−0.86	−0.82	−0.48
$AR(2)$-MRA	2.52	0.10	4.61	1.41	5.47	5.78	9.39	12.67	16.40	18.84	25.09	26.58	31.71	34.22	17.45
$AR(3)$-MRA	3.38	1.54	5.08	3.06	6.94	6.45	10.79	13.18	18.43	19.60	25.70	28.35	30.75	34.36	18.67
$AR(6)$-MRA	8.08	3.93	8.28	3.46	9.84	17.10	18.63	23.81	29.56	33.25	40.73	38.43	41.38	42.86	30.47
$ARMA(1,1)$-MRA	3.05	1.15	4.05	2.75	5.91	6.07	9.03	9.58	12.25	14.20	18.44	19.76	22.60	24.69	12.06
$ARMA.auto$-MRA	0.31	0.12	1.35	0.46	−0.00	−0.08	1.40	1.08	1.20	1.84	0.86	2.41	1.28	1.83	1.00
50-day rolling window forecast															
$AR(1)$-MRA	−0.62	−1.36	−1.08	−0.98	0.14	−0.22	0.03	0.03	−0.05	0.24	−0.68	−0.58	−0.40	−0.11	−0.40
$AR(2)$-MRA	−0.19	−0.35	0.56	−0.57	0.09	0.43	0.69	0.86	0.82	0.39	1.33	0.59	0.90	0.22	0.41
$AR(3)$-MRA	−0.54	0.17	2.07	−0.96	0.79	0.28	1.23	1.72	1.16	0.99	3.13	1.75	1.79	0.98	1.05
$AR(6)$-MRA	0.85	0.62	3.14	0.30	3.75	4.89	5.83	8.29	11.04	12.57	17.72	17.90	22.12	24.30	10.54
$ARMA(1,1)$-MRA	−1.55	0.48	−0.10	1.85	1.82	2.86	3.34	3.32	5.54	6.24	7.49	7.50	9.25	10.56	4.35
$ARMA.auto$-MRA	−1.89	−1.37	−0.42	−1.61	−1.34	−0.69	0.36	−0.13	0.69	0.06	0.90	−0.26	−0.33	−0.19	−0.45

Table 1.3: **Forecast Improvement (MAE) for S&P500 Index Return with Different Horizons: *t*-statistics**

The numbers in the table are the Paired t-statistics of the difference in means between the absolute prediction errors from the benchmark models and corresponding multiresolution-analysis augmented forecasts. Multiresolution analysis and Rolling window forecasts are conducted in the same way as described in Table 1.1. The sample is based on the daily returns of the S&P500 index from February 14, 2011 to August 19, 2016.

	Forecast horizon, daily basis													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
10-day rolling window forecast														
$AR(1)$-MRA	2.69	−2.33	0.54	−0.15	0.50	0.91	1.05	1.04	1.15	1.10	1.15	1.13	1.13	1.12
$AR(2)$-MRA	6.44	1.72	2.44	1.32	1.46	0.96	1.04	1.02	0.99	0.99	1.01	1.00	0.99	1.00
$AR(3)$-MRA	10.44	4.10	4.04	2.28	1.58	1.14	0.93	0.68	0.54	0.29	0.10	−0.14	−0.36	−0.53
$ARMA(1,1)$-MRA	6.03	3.42	4.52	2.90	3.46	2.77	3.12	3.04	2.78	2.70	2.54	2.48	2.34	2.21
$ARMA.auto$-MRA	−1.87	−1.73	−0.61	−2.60	−2.68	−1.51	−1.77	−1.08	−2.13	−1.82	−1.70	−1.35	−2.51	−2.40
30-day rolling window forecast														
$AR(1)$-MRA	−0.33	−2.11	−2.11	−1.81	−1.06	−1.51	−0.30	−1.50	−0.60	−0.25	−1.22	−1.72	−1.16	−0.64
$AR(2)$-MRA	0.41	−1.51	0.04	0.15	1.07	0.81	1.28	0.94	0.94	1.20	1.43	0.92	0.86	0.98
$AR(3)$-MRA	1.33	0.85	1.71	0.78	2.23	0.60	0.69	0.47	0.78	0.83	1.22	1.10	0.89	0.83
$AR(6)$-MRA	4.11	2.96	1.87	0.82	1.74	1.56	1.92	1.15	1.66	1.48	1.74	1.75	1.26	1.24
$ARMA(1,1)$-MRA	0.78	0.75	0.71	0.63	2.44	2.87	2.71	2.38	2.19	2.12	2.76	2.55	2.51	2.20
$ARMA.auto$-MRA	−2.51	−0.57	−1.11	−1.47	−1.09	−0.98	0.72	0.43	0.70	0.83	1.36	1.89	0.37	1.06
50-day rolling window forecast														
$AR(1)$-MRA	−2.14	−2.11	−2.16	−1.93	−1.27	−1.16	−1.45	−1.10	−0.52	−0.39	−1.13	−1.22	−1.27	−0.65
$AR(2)$-MRA	−2.04	−0.93	−0.52	−0.47	−0.86	0.01	1.00	0.72	0.78	0.72	1.24	−0.21	0.53	1.34
$AR(3)$-MRA	−2.29	−0.60	1.29	−1.80	−0.93	−0.57	0.75	0.61	−0.83	0.13	1.19	0.44	0.12	0.68
$AR(6)$-MRA	−0.65	0.11	0.54	−0.96	−0.07	0.65	0.67	0.63	1.41	1.16	1.46	1.01	1.14	1.26
$ARMA(1,1)$-MRA	−2.63	0.64	−0.81	2.51	1.52	0.58	1.04	1.77	2.66	1.60	1.20	1.76	1.36	1.87
$ARMA.auto$-MRA	−3.92	−3.82	−2.50	−3.30	−3.73	−2.12	−0.78	−1.07	1.48	−1.51	0.25	−1.72	−1.73	−2.41

Table 1.4: **Forecast Improvement (rMSE) for S&P500 Index Return with Different Horizons: t-statistics**

The numbers in the table are the Paired t-statistics of the difference in means between the squared prediction errors from the benchmark models and corresponding multiresolution-analysis augmented forecasts. Multiresolution analysis and Rolling window forecasts are conducted in the same way as described in Table 1.2. The sample is based on the daily returns of the S&P500 index from February 14, 2011 to August 19, 2016.

	Forecast horizon, daily basis													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
10-day rolling window forecast														
$AR(1)$-MRA	2.95	0.15	1.24	0.97	1.04	1.05	1.05	1.05	1.04	1.03	1.03	1.03	1.02	1.02
$AR(2)$-MRA	4.44	1.63	1.26	1.08	1.04	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$AR(3)$-MRA	6.25	2.61	1.81	1.47	1.39	1.36	1.35	1.33	1.27	1.16	0.96	0.63	0.18	−0.27
$ARMA(1,1)$-MRA	5.49	2.74	2.23	1.98	1.84	1.96	2.05	2.03	1.98	1.92	1.82	1.69	1.57	1.46
$ARMA.auto$-MRA	0.20	0.51	0.38	−0.49	−0.98	0.68	0.28	0.91	0.08	0.79	−0.03	0.40	0.39	0.10
30-day rolling window forecast														
$AR(1)$-MRA	0.70	−0.89	−0.83	−1.30	0.29	−0.78	0.64	−0.47	−0.06	0.10	−1.17	−1.57	−0.91	−0.90
$AR(2)$-MRA	1.84	0.10	1.28	0.75	1.04	0.97	1.17	1.00	1.01	1.00	1.04	0.97	1.00	0.98
$AR(3)$-MRA	2.26	1.40	1.80	1.24	1.61	1.06	1.34	1.14	1.13	1.09	1.14	1.05	1.06	1.02
$AR(6)$-MRA	4.17	2.17	1.93	1.02	1.46	1.59	1.50	1.24	1.15	1.38	1.38	1.28	1.07	1.09
$ARMA(1,1)$-MRA	1.62	1.03	1.78	1.71	2.20	2.03	2.23	2.02	2.00	2.03	2.05	2.01	1.97	2.02
$ARMA.auto$-MRA	0.16	0.10	0.88	0.41	−0.01	−0.11	1.83	1.37	1.47	1.48	1.34	1.96	1.60	1.54
50-day rolling window forecast														
$AR(1)$-MRA	−0.78	−1.07	−1.96	−1.31	0.30	−0.39	0.06	0.05	−0.13	0.45	−1.59	−1.25	−1.05	−0.29
$AR(2)$-MRA	−0.18	−0.44	0.49	−0.73	0.11	0.62	1.32	1.01	0.99	1.07	1.31	0.71	0.92	1.14
$AR(3)$-MRA	−0.44	0.26	1.75	−1.28	0.80	0.31	1.48	1.20	0.81	1.03	1.55	0.80	0.89	0.62
$AR(6)$-MRA	0.54	0.73	1.49	0.16	1.22	1.28	1.36	1.15	1.23	1.38	1.29	1.12	1.12	1.16
$ARMA(1,1)$-MRA	−1.20	0.56	−0.09	1.60	2.11	1.28	1.29	1.62	1.33	1.27	1.25	1.43	1.18	1.18
$ARMA.auto$-MRA	−1.50	−1.56	−0.49	−1.94	−2.21	−1.26	0.67	−0.34	1.42	0.12	1.71	−0.67	−0.74	−0.55

Table 1.5: **Forecast Improvement (MAE) for S&P500 Net Total Return with Different Horizons**

The numbers in the table are percentage improvements (decreases) in mean absolute error (MAE) of multi-step forecasts by multiresolution-analysis augmented method, comparing with forecasts using corresponding benchmark models without using multiresolution analysis. Multiresolution analysis is conducted by level-2 MODWT with haar wavelet. The sample is based on the daily returns of the S&P500 Net Total Return from February 14, 2011 to August 19, 2016. (The "net total return" version of the index reflects the effects of both dividend reinvestment and withholding tax deduction.) Rolling window forecasts are conducted with 10, 30, and 50 days as the window length. Each forecast practice is based on the information limited to the observations during corresponding window period. For example, the procedure for an h-step 10-day rolling window forecast includes: 10-day observations are used to generate wavelet-based hierarchical multiple time series; an h-step out-of-sample forecast is conducted for each times series without re-estimation, where multi-step forecasts, $\hat{y}_{T+h}$ with $h > 1$, take previous period's predicted values as input; refined combination forecasts are calculated based on $\hat{y}_{T+h}$ at all levels. The last column are the forecast improvements based on the average MAE among all h-step forecasts. The *ARMA.auto* is based on an optimal ARIMA model using the automatic algorithm of Hyndman and Khandakar (2008).

	Forecast horizon, daily basis														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave
10-day rolling window forecast															
$AR(1)$-MRA	2.02	−1.43	0.44	−0.04	1.00	2.50	4.49	6.89	11.24	15.52	21.32	26.70	31.95	35.86	18.80
$AR(2)$-MRA	7.46	2.19	6.30	6.09	14.18	16.86	27.65	34.46	39.74	42.48	44.26	44.98	45.43	45.71	45.15
$AR(3)$-MRA	13.60	6.79	12.76	11.95	15.42	16.92	18.98	17.30	16.72	11.56	7.45	−0.21	−10.10	−22.55	−15.91
$ARMA(1,1)$-MRA	5.49	2.63	4.25	2.92	4.86	4.66	6.70	8.23	10.25	13.12	15.06	18.69	22.34	25.10	12.68
$ARMA.auto$-MRA	−1.51	−1.17	−0.33	−1.78	−2.08	−1.08	−1.31	−0.87	−1.91	−1.59	−1.70	−1.40	−2.96	−2.85	−1.61
30-day rolling window forecast															
$AR(1)$-MRA	−0.24	−1.62	−1.08	−1.13	−0.41	−0.73	−0.12	−0.80	−0.26	−0.13	−0.54	−0.94	−0.53	−0.32	−0.63
$AR(2)$-MRA	0.43	−1.01	0.04	0.11	0.92	0.64	1.34	1.12	1.32	2.09	3.04	2.29	2.63	3.57	1.35
$AR(3)$-MRA	1.56	0.53	1.26	0.62	1.87	0.53	0.78	0.60	1.29	1.54	2.70	3.16	2.86	3.34	1.64
$AR(6)$-MRA	5.76	2.34	2.24	0.79	2.34	2.95	3.85	2.70	4.75	5.07	8.22	8.90	7.96	9.37	4.98
$ARMA(1,1)$-MRA	0.89	0.51	0.51	0.19	1.75	2.01	2.53	2.44	2.59	2.55	4.06	3.87	4.31	4.11	2.33
$ARMA.auto$-MRA	−2.81	−0.51	−0.91	−0.99	−0.59	−0.50	0.38	0.28	0.45	0.31	0.69	0.74	0.18	0.53	−0.20
50-day rolling window forecast															
$AR(1)$-MRA	−1.63	−1.71	−1.03	−1.10	−0.41	−0.47	−0.47	−0.44	−0.16	−0.13	−0.31	−0.40	−0.34	−0.17	−0.63
$AR(2)$-MRA	−1.95	−0.57	−0.34	−0.23	−0.43	0.01	0.30	0.21	0.24	0.14	0.39	−0.05	0.14	0.28	−0.14
$AR(3)$-MRA	−2.60	−0.30	0.69	−0.81	−0.45	−0.23	0.32	0.24	−0.39	0.04	0.57	0.21	0.03	0.35	−0.17
$AR(6)$-MRA	−0.83	0.07	0.37	−0.70	−0.05	0.62	0.74	0.78	1.77	1.66	2.28	1.81	2.22	2.77	0.95
$ARMA(1,1)$-MRA	−3.15	0.37	−0.46	1.14	0.90	0.22	0.61	1.16	1.78	1.22	1.07	1.62	1.14	2.04	0.69
$ARMA.auto$-MRA	−4.88	−2.49	−1.67	−1.69	−1.98	−0.97	−0.45	−0.41	0.53	−0.66	−0.00	−0.55	−0.50	−0.75	−1.18

Table 1.6: **Forecast Improvement (rMSE) for S&P500 Net Total Return with Different Horizons**

The numbers in the table are percentage improvements (decreases) in root mean squared error (rMSE) of multi-step forecasts by multiresolution-analysis augmented method, comparing with forecasts using corresponding benchmark models without using multiresolution analysis. Multiresolution analysis is conducted by level-2 MODWT with haar wavelet. The sample is based on the daily returns of the S&P500 Net Total Return from February 14, 2011 to August 19, 2016. (The “net total return” version of the index reflects the effects of both dividend reinvestment and withholding tax deduction.) Rolling window forecasts are conducted with 10, 30, and 50 days as the window length. Each forecast practice is based on the information limited to the observations during corresponding window period. For example, the procedure for an h-step 10-day rolling window forecast includes: 10-day observations are used to generate wavelet-based hierarchical multiple time series; an h-step out-of-sample forecast is conducted for each times series without re-estimation, where multi-step forecasts, $\hat{y}_{T+h}$ with $h > 1$, take previous period's predicted values as input; refined combination forecasts are calculated based on $\hat{y}_{T+h}$ at all levels. The last column are the forecast improvements based on the average rMSE among all h-step forecasts. The $ARMA.auto$ is based on an optimal ARIMA model using the automatic algorithm of Hyndman and Khandakar (2008).

	Forecast horizon, daily basis														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave
10-day rolling window forecast															
$AR(1)$-MRA	3.64	0.36	4.40	9.25	19.02	28.29	35.94	39.65	42.13	43.17	43.99	44.46	44.86	45.12	43.39
$AR(2)$-MRA	11.01	9.96	28.23	37.51	45.52	45.29	46.28	46.02	46.19	46.13	46.16	46.15	46.15	46.15	46.15
$AR(3)$-MRA	17.56	17.42	35.44	37.62	42.70	40.99	40.78	39.75	38.52	36.41	33.22	28.23	20.53	9.02	14.54
$ARMA(1,1)$-MRA	5.88	3.59	9.07	9.01	16.21	23.39	30.45	36.36	40.51	43.73	44.34	45.28	45.78	45.94	41.54
$ARMA.auto$-MRA	0.32	0.54	0.54	−0.54	−1.44	0.81	0.65	2.11	0.33	2.99	−0.05	1.76	1.84	0.63	0.76
30-day rolling window forecast															
$AR(1)$-MRA	0.70	−1.12	−0.54	−1.11	0.15	−0.52	0.44	−0.40	−0.04	0.09	−0.98	−1.64	−0.85	−0.82	−0.47
$AR(2)$-MRA	2.53	0.09	4.56	1.39	5.39	5.67	9.24	12.46	16.13	18.51	24.78	26.23	31.40	33.89	17.17
$AR(3)$-MRA	3.39	1.54	5.01	3.03	6.87	6.32	10.63	12.96	18.18	19.32	25.38	28.06	30.48	34.04	18.41
$AR(6)$-MRA	8.10	3.94	8.19	3.41	9.78	17.02	18.59	23.60	29.40	33.14	40.60	38.43	41.22	42.77	30.32
$ARMA(1,1)$-MRA	2.82	1.18	3.66	2.39	5.20	5.55	8.13	8.95	11.13	12.82	16.94	18.04	20.74	22.74	10.90
$ARMA.auto$-MRA	0.30	−0.02	1.03	0.51	0.06	−0.03	1.29	1.10	1.22	1.74	1.03	2.30	1.19	1.87	0.97
50-day rolling window forecast															
$AR(1)$-MRA	−0.62	−1.37	−1.07	−0.97	0.14	−0.22	0.03	0.03	−0.05	0.24	−0.68	−0.58	−0.40	−0.11	−0.40
$AR(2)$-MRA	−0.17	−0.35	0.55	−0.58	0.07	0.41	0.67	0.83	0.79	0.38	1.30	0.56	0.86	0.20	0.39
$AR(3)$-MRA	−0.51	0.17	2.05	−0.96	0.76	0.25	1.20	1.68	1.11	0.96	3.06	1.69	1.71	0.91	1.01
$AR(6)$-MRA	0.86	0.64	3.09	0.35	3.68	4.93	5.70	8.32	10.93	12.62	17.58	17.93	21.98	24.28	10.50
$ARMA(1,1)$-MRA	−1.56	0.44	−0.06	1.81	1.83	2.80	3.35	3.36	5.47	6.23	7.48	7.46	9.22	10.55	4.34
$ARMA.auto$-MRA	−1.87	−1.49	−0.57	−1.58	−1.27	−0.78	0.30	−0.16	0.67	0.05	0.85	−0.25	−0.32	−0.14	−0.47

Table 1.7: **Forecast Improvement (MAE) for S&P500 Net Total Return with Different Horizons: *t*-statistics**

The numbers in the table are the Paired t-statistics of the difference in means between the absolute prediction errors from the benchmark models and corresponding multiresolution-analysis augmented forecasts. Multiresolution analysis and Rolling window forecasts are conducted in the same way as described in Table 1.5. The sample is based on the daily returns of the S&P500 Net Total Return from February 14, 2011 to August 19, 2016. (The "net total return" version of the index reflects the effects of both dividend reinvestment and withholding tax deduction.)

	Forecast horizon, daily basis													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
10-day rolling window forecast														
$AR(1)$-MRA	2.69	−2.31	0.58	−0.03	0.61	0.96	1.07	1.06	1.17	1.12	1.16	1.15	1.15	1.14
$AR(2)$-MRA	6.47	1.72	2.44	1.31	1.47	0.97	1.04	1.02	0.99	0.99	1.01	1.00	0.99	1.00
$AR(3)$-MRA	10.45	3.98	3.90	2.24	1.50	1.08	0.89	0.65	0.55	0.34	0.20	−0.00	−0.21	−0.39
$ARMA(1,1)$-MRA	5.89	3.20	4.41	2.44	3.19	2.35	2.60	2.37	2.22	2.14	1.86	1.79	1.68	1.56
$ARMA.auto$-MRA	−1.79	−1.63	−0.43	−2.54	−2.67	−1.44	−1.58	−1.05	−2.02	−1.72	−1.67	−1.27	−2.57	−2.39
30-day rolling window forecast														
$AR(1)$-MRA	−0.30	−2.12	−2.11	−1.81	−1.07	−1.51	−0.28	−1.49	−0.61	−0.25	−1.24	−1.73	−1.18	−0.61
$AR(2)$-MRA	0.42	−1.50	0.05	0.17	1.06	0.81	1.29	0.94	0.92	1.20	1.45	0.92	0.85	0.98
$AR(3)$-MRA	1.33	0.86	1.68	0.82	2.25	0.60	0.68	0.46	0.77	0.82	1.21	1.10	0.88	0.82
$AR(6)$-MRA	4.11	2.96	1.86	0.77	1.72	1.55	1.97	1.17	1.68	1.47	1.75	1.77	1.29	1.26
$ARMA(1,1)$-MRA	0.76	0.72	0.67	0.30	2.20	2.61	2.68	2.47	2.24	2.05	2.90	2.52	2.48	2.15
$ARMA.auto$-MRA	−2.46	−0.67	−1.23	−1.49	−0.98	−0.91	0.76	0.50	0.96	0.58	1.50	1.46	0.44	1.15
50-day rolling window forecast														
$AR(1)$-MRA	−2.14	−2.13	−2.17	−1.91	−1.27	−1.16	−1.42	−1.08	−0.54	−0.37	−1.12	−1.25	−1.29	−0.65
$AR(2)$-MRA	−2.04	−0.92	−0.55	−0.46	−0.85	0.02	1.00	0.70	0.76	0.70	1.25	−0.19	0.49	1.34
$AR(3)$-MRA	−2.30	−0.58	1.25	−1.77	−0.93	−0.59	0.74	0.57	−0.91	0.10	1.17	0.41	0.07	0.64
$AR(6)$-MRA	−0.62	0.12	0.48	−0.95	−0.06	0.65	0.67	0.70	1.38	1.21	1.43	1.05	1.12	1.27
$ARMA(1,1)$-MRA	−2.70	0.57	−0.71	2.31	1.61	0.40	0.97	1.76	2.50	1.60	1.24	1.73	1.12	1.82
$ARMA.auto$-MRA	−3.97	−3.92	−2.67	−3.09	−3.50	−2.14	−0.97	−1.02	1.32	−1.81	−0.01	−1.63	−1.38	−2.30

Table 1.8: **Forecast Improvement (rMSE) for S&P500 Net Total Return with Different Horizons: *t*-statistics**

The numbers in the table are the Paired t-statistics of the difference in means between the squared prediction errors from the benchmark models and corresponding multiresolution-analysis augmented forecasts. Multiresolution analysis and Rolling window forecasts are conducted in the same way as described in Table 1.6. The sample is based on the daily returns of the S&P500 index from February 14, 2011 to August 19, 2016. (The "net total return" version of the index reflects the effects of both dividend reinvestment and withholding tax deduction.)

	Forecast horizon, daily basis													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
10-day rolling window forecast														
$AR(1)$-MRA	2.94	0.20	1.24	0.99	1.05	1.06	1.06	1.05	1.05	1.04	1.04	1.03	1.03	1.03
$AR(2)$-MRA	4.44	1.63	1.26	1.08	1.05	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$AR(3)$-MRA	6.26	2.55	1.76	1.43	1.35	1.32	1.31	1.30	1.26	1.20	1.08	0.90	0.62	0.24
$ARMA(1,1)$-MRA	5.25	2.66	2.25	1.89	1.69	1.78	1.82	1.77	1.72	1.62	1.53	1.42	1.32	1.24
$ARMA.auto$-MRA	0.21	0.57	0.39	−0.44	−0.96	0.65	0.27	0.90	0.14	0.80	−0.01	0.41	0.40	0.13
30-day rolling window forecast														
$AR(1)$-MRA	0.70	−0.90	−0.82	−1.30	0.30	−0.78	0.65	−0.46	−0.06	0.11	−1.18	−1.56	−0.90	−0.90
$AR(2)$-MRA	1.85	0.09	1.27	0.75	1.04	0.97	1.17	1.00	1.02	1.00	1.04	0.97	1.00	0.98
$AR(3)$-MRA	2.28	1.40	1.80	1.24	1.62	1.06	1.34	1.14	1.14	1.09	1.14	1.05	1.06	1.02
$AR(6)$-MRA	4.18	2.17	1.92	1.01	1.47	1.59	1.51	1.24	1.16	1.39	1.39	1.29	1.07	1.10
$ARMA(1,1)$-MRA	1.58	1.06	1.80	1.54	2.22	1.92	2.18	1.93	1.90	1.91	1.99	1.90	1.84	1.92
$ARMA.auto$-MRA	0.16	−0.02	0.72	0.47	0.08	−0.04	1.69	1.40	1.49	1.43	1.59	1.88	1.53	1.59
50-day rolling window forecast														
$AR(1)$-MRA	−0.77	−1.08	−1.95	−1.30	0.31	−0.39	0.06	0.04	−0.14	0.45	−1.58	−1.24	−1.05	−0.28
$AR(2)$-MRA	−0.16	−0.45	0.48	−0.74	0.09	0.61	1.31	1.00	0.98	1.06	1.31	0.70	0.91	1.11
$AR(3)$-MRA	−0.42	0.27	1.75	−1.29	0.79	0.28	1.47	1.20	0.79	1.01	1.56	0.79	0.87	0.59
$AR(6)$-MRA	0.55	0.75	1.47	0.19	1.20	1.31	1.34	1.16	1.22	1.40	1.29	1.13	1.12	1.16
$ARMA(1,1)$-MRA	−1.20	0.52	−0.06	1.58	2.13	1.25	1.30	1.65	1.31	1.27	1.24	1.43	1.17	1.18
$ARMA.auto$-MRA	−1.49	−1.69	−0.67	−1.90	−2.11	−1.41	0.56	−0.42	1.38	0.11	1.65	−0.66	−0.73	−0.41

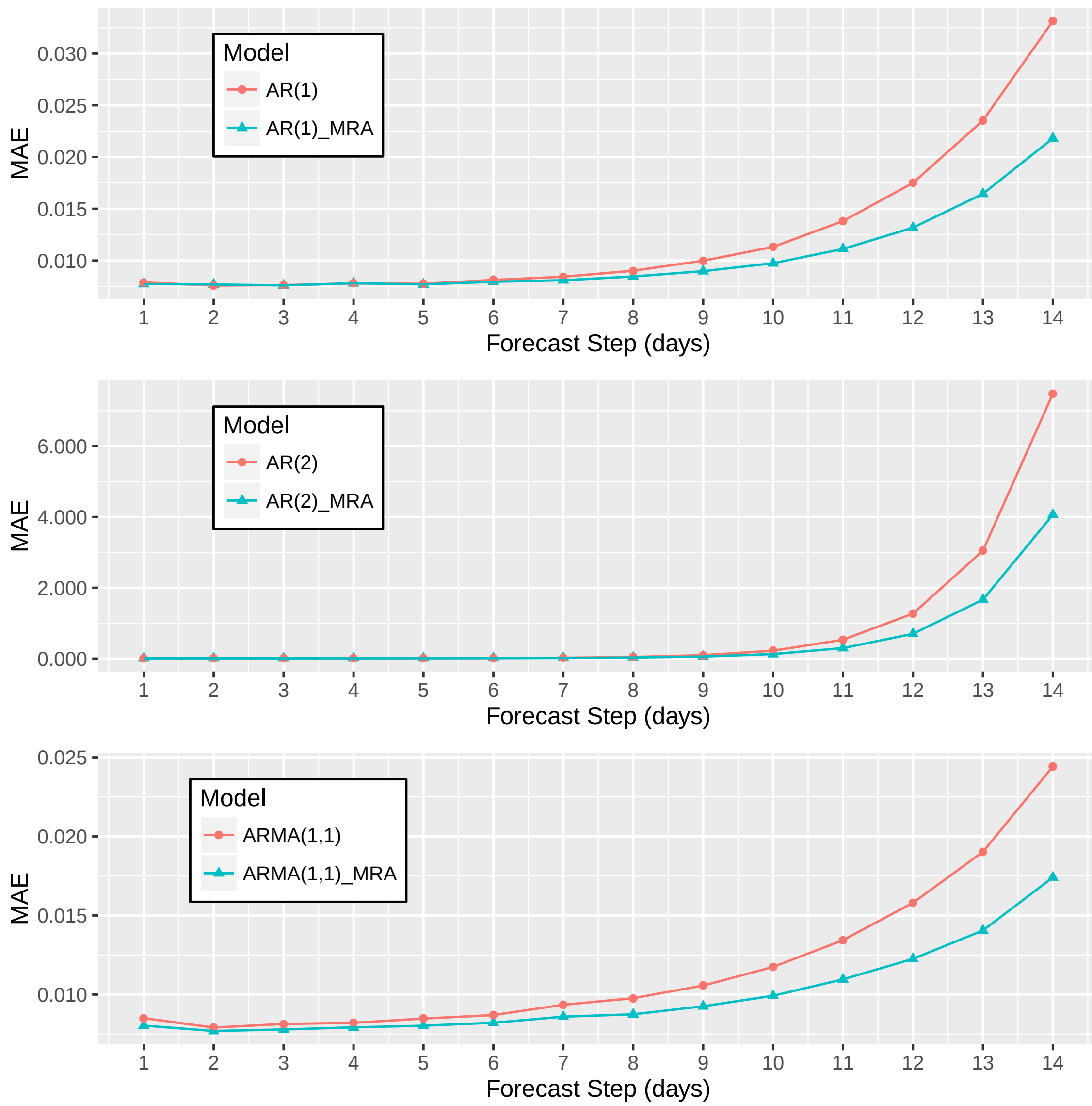


Figure 1.2: **Forecast Improvement (MAE) with Different Horizons**
The results correspond to the ones in the panel of 10-day rolling window in table 1.1.

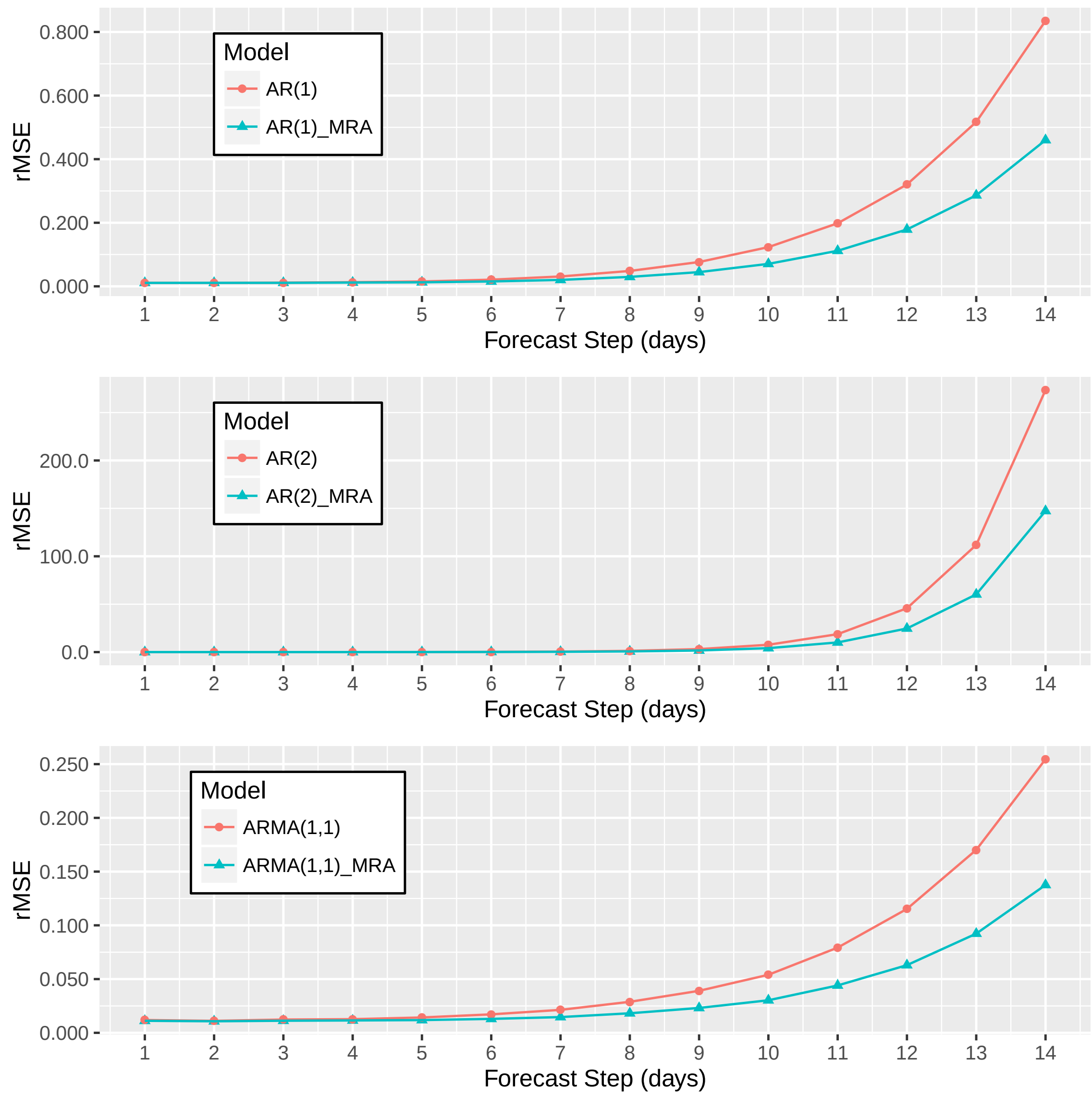


Figure 1.3: **Forecast Improvement (rMSE) with Different Horizons**
The results correspond to the ones in the panel of 10-day rolling window in table 1.2.

Chapter 2

Application of Network Structures in Stock Return Volatility Forecasting

2.1 Introduction

Equity market volatility is broadly understood as the degree of variation in stock prices or returns. Compared to the fairly large econometric literature focusing on pure time-series modeling and forecasting of volatility (Andersen, Bollerslev, Christoffersen, and Diebold (2006) provide a survey on these topics), research on economic channels of volatility is relatively scarce. As a seminal paper on the drivers of equity market volatility, Schwert (1989) investigates the relationship of stock volatility with macroeconomic volatility, economic activity, financial leverage and stock trading activity. More recently, adopting a comprehensive approach, Paye (2012) and Christiansen, Schmeling, and Schrimpf (2012) model volatility in a predictive regression setting. Motivated by existing theoretical literature, Paye (2012) identifies a set of candidate predictors and tests the ability of these variables to improve volatility forecasts. Christiansen et al. (2012) perform an even more comprehensive examination of financial volatility in the sense that they investigate a larger set of potential predictors and study not only equity market but also volatility in other asset classes (i.e., bonds, foreign exchange and commodities).

Note that the majority of the existing literature has focused exclusively on aggregate equity market volatility. In contrast, the present paper investigates stock return volatility from the perspective of individual firms. More important, we focus on one particular channel of volatility—the transfer of volatility along supply chains.

Today, no firm is an isolated island—firms are connected to one another through different types of linkages. Some of these links are direct and explicit, while others are relatively obscure. Allen and Babus (2008) provide a survey of possible sources of connections between financial institutions and how these connections are modeled and explored to answer important economic questions. In this study, we focus on the customer-supplier links between firms—these links are clearly defined, contractual, and founded on real trading activities.

There are interesting studies investigating customer-supplier relationships. For example, Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012) study the systemic risks originating from intersectoral input-output linkages and argue that sizable aggregate fluctuations may originate from microeconomic idiosyncratic shocks only if there are significant asymmetries in the roles that sectors play as suppliers to others. In another examination of the linkages along the supply chain, Hertzel, Li, Officer, and Rodgers (2008) study the wealth effects of distress and bankruptcy filing for the suppliers and customers of filing firms and find that significant contagion effects extend to suppliers of the filing firms but not the customers. In a recent paper, Gençay, Signori, Xue, Yu, and Zhang (2015) introduce an econometric network model to appropriately analyze counterparty risks and find that for each supplier, customers' leverage and option implied volatilities are significant determinants of corporate credit spreads in the period after the 2008 to 2009 U.S. recession.

Shocks can propagate through supply chains. Customers are crucial to suppliers because they play an indispensable role in fulfilling most firms' ultimate goal—selling goods and/or services for a profit. Cohen and Frazzini (2008) show that firms' real operations, measured by sales and operating income, are significantly more correlated when they are linked through customer-supplier relationships, which affirms the intuition that there are significant comovements in the underlying cash flows of the linked firms. Furthermore, as documented in Hertzel et al. (2008), when a firm experiences financial distress, which largely reflects a shift in demand away from the firm, this may also reduce the derived demand for its suppliers' output—this is a typical example of how shocks propagate through supply chains. As firms are stakeholders in their customers' operations—the financial health of their major customers affects their own profitability, and shocks to the customers have resulting effects on the supplier firm—it is natural to hypothesize that the customer-supplier link is an important channel for the propagation of stock return volatility. Specifically, customers' return volatility predicts that of suppliers.

This paper tests and examines this hypothesis in a predictive regression setting. As a seminal paper on the drivers of equity market volatility, Schwert (1989) provides evidence that there is a strong association between stock return volatility and stock trading activity. Our analysis shows that the effect of customer volatility is approximately 10 times as large as trading value on supplier's volatility. Our findings are robust to controlling for variables capturing the time-series properties of stock return volatility and a set of idiosyncratic, industry and market factors. Further, they are tested under various assumptions regarding the activeness of customer-supplier linkages and using different estimation methods. In particular, we explicitly address the potential concern of endogeneity and confirm our results through estimations with instrumental variables (IVs) using the two-step efficient Generalized Method of Moments (GMM).

Moreover, using various benchmark models and rolling estimation windows, our out-of-sample tests show that incorporating the customer channel improves forecasts of a supplier's volatility. In particular, even in the most aggressive case, namely, using a benchmark model with a large set of forecasting variables that are documented to be important predictors of volatility, our results still provide evidence of improvements in forecasting when the customer channel is included.

In addition to demonstrating the predictive power of customer volatility, we test how the public's awareness of customer-supplier linkages affects the intensity of volatility transfer. Specifically, we investigate the interaction between customer volatility and analyst coverage. As the research activities and recommendations of financial analysts reveal a firm's information to the market, higher analyst coverage implies that the public has better access to information regarding a firm's operations and financial conditions and is more likely to be aware of the identities of its major customers. As expected, our results show that the transfer of volatility from customers to suppliers is more pronounced for firms with higher analyst coverage (after controlling for the size of the firm, measured by market capitalization and trading value). Our interpretation is that when investors are more aware of a firm's principal customers, they anticipate the propagation of shocks through these related firms and respond more actively to news on firm's customers; that is, they incorporate such news into their investment and asset allocation decisions regarding the supplier companies. This trend contributes to the stronger association between customer and supplier volatility.

We would like to emphasize that our analysis is based on a large sample over a long period in the equity market. We employ panel data spanning approximately forty years, from 1977 to 2015. After matching with data on the control variables we consider, our final sample contains 2,738 unique suppliers with a total of 134,007 monthly observations.

The remainder of the paper is organized as follows. Section 2.2 introduces the construction of customer-supplier networks and how we model stock return volatility in a predictive regression setting. Section 2.3 describes the data. Section 2.4 reports the estimation results. Section 2.5 considers five robustness checks: we control for industry effects; examine alternative assumptions regarding the activeness of customer-supplier linkages; address the potential correlation between individual fixed effects and the autoregressive term; address the potential concern of endogeneity using IV estimation; and for comparison, show that while customers' return volatility predicts that of suppliers, return volatilities of randomly selected firms are not significant determinants of a firms's volatility. The out-of-sample analysis is performed in Section 2.6. Section 2.7 examines the interaction between customer volatility and analyst coverage. Section 2.8 concludes.

2.2 Methodology

2.2.1 Customer-Supplier Networks and Adjacency Matrices

In a survey paper, Allen and Babus (2008) report that networks, which are generally understood as collections of nodes and links between nodes, can be useful representations of economic or financial systems. Nodes represent entities in the system; links describe certain relationships between the entities.

In this paper, we examine customer-supplier networks, where each firm is a "node", and a customer-supplier relationship is a "link" between two firms. The structure of the network can be characterized by an adjacency matrix, G, which is a square matrix with dimension of the number of nodes (i.e., firms) in the network. The entry in the ith row and jth column of G, $(G)_{ij}$, is one if and only if i (j) is the supplier (customer) of j (i), zero otherwise.

[Insert Figure 2.1 about here.]

Consider the simple network depicted in Figure 2.1, v_i, $i = 1, ..., 5$, denotes the firm; the arrow indicates the flow of output. For example, the arrow between v_1 and v_2 indicates that firm 1 (2) is the supplier (customer) of firm 2 (1). Matrix G characterizing the structure of this network is therefore

$$G = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The second row of G, for example, refers to firm 2, which indicates that firm 2 has only one customer, which is firm 4, as only the fourth entry is one. More generally, the ith row of G captures firm i's first-order (i.e., immediate) customer linkages.

The adjacency matrix G we have referred to thus far is unweighted, in the sense that it has entries of either one or zero. In some applications, it is useful to introduce the concept of the *strength* of a link. In this paper, we use a sales-weighted matrix G to capture the relative importance of customers.[8] First, we construct an unweighted G. Next, for each supplier (i.e., each row) in G, links (i.e., entries that have a value of one) are weighted by the amount of sales made to the target customer, normalized by the observed total amount of sales (i.e., the sum of all sales to customers) of this supplier in this period. The sum of the entries in each row of the sales-weighted G is equal to one. Using this weighting, from a supplier's perspective, greater importance is assigned to customers that account for a larger shares of trades.[9]

2.2.2 Stock Return Volatility

Realized stock return volatility is often measured by the realized variance or the square root of the realized variance of the (excess) stock returns. (For example, see Schwert (1989), Andersen et al. (2006), Corsi (2009), Paye (2012), Christiansen et al. (2012) and Cao and Han (2013).) Following the same approach, we focus on modeling and forecasting the sample standard deviation of daily stock returns over one month. Specifically, the realized stock return volatility of firm i in month t, $RV_{i,t}$, is characterized by

$$RV_{i,t} = \sqrt{\frac{1}{T_{i,t}-1}\sum_{\tau=1}^{T_{i,t}}(DailyRet_{i,t,\tau} - \overline{DailyRet_{i,t}})^2}, \tag{2.1}$$

where, for firm i, $T_{i,t}$ is the number of trading days observed in month t; $DailyRet_{i,t,\tau}$ is the daily return on the τth trading day of month t (i.e., $DailyRet_{i,t,\tau} = \frac{P_{i,t,\tau}-P_{i,t,\tau-1}}{P_{i,t,\tau-1}}$, where $P_{i,t,\tau}$ is the closing stock price on the τth trading day of month t); and $\overline{DailyRet_{i,t}}$ is the sample average of daily returns in month t.

For each firm i, its customers' stock return volatility in month t is measured and denoted by $(G\cdot RV)_{i,t} = (G_t \cdot RV_t)_i$. Suppose that there are n firms in the customer-supplier network in month t, G_t is the $n \times n$ sales-weighted adjacency matrix in month t, RV_t is the $n \times 1$ vector containing each firm's return volatility in month t, and $(G_t \cdot RV_t)$ is therefore a vector capturing the return volatility of each firm's customers in month t. Specifically, the ith entry in $(G_t \cdot RV_t)$ is the sales-weighted average of the return volatilities of firm i's customers in month t.

We model volatility in a predictive regression setting—all of the right-hand-side variables are one month prior to the dependent variable, especially focusing on the channel from customer volatility to that of the supplier:

$$RV_{i,t} = \beta_0 + \beta_1 \left(G \cdot RV\right)_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t} \tag{2.2}$$

where $ControlVariables$ is a row vector containing a set of time-series, market and idiosyncratic factors that are introduced in the next section, and Γ is a column vector of coefficients.[10]

As also noted by Cohen and Frazzini (2008), current U.S. financial accounting regulation requires public firms to report the customers that account for at least 10% of their total yearly sales (but not their suppliers); thus, our data source provides more information about firms' major customers (but not major suppliers).[11] Therefore, we focus our investigation on the effects of customers, that is, the transfer of volatility from customers to suppliers.

2.2.3 Control Variables

Motivated by the existing theoretical and empirical literature, we incorporate a set of control variables into our analysis. First, to capture the time-series properties of stock return volatility, we include an autoregressive term, the one-month-lagged return volatility (RV^m), in our model. In addition, we apply the Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) from Corsi (2009) and Andersen, Bollerslev, and Diebold (2007), including return volatilities over different time horizons other than one month: one quarter (RV^q), one-half year (RV^{hy}), and one year (RV^y). They are defined and calculated using daily stock returns in the same way as RV in Equation 2.1, except that they are over longer time horizons.

To control for market-level volatility, we include the sample standard deviation of daily returns on the Standard & Poor's 500 Composite Index over one month, $RV_{S\&P}$. In addition, we include *yield curve slope* and *Baa-Aaa spread*. The slope of the yield curve is measured as the difference between the 10-year and 2-year Treasury Constant Maturity rates, $r^{10} - r^2$; the Baa-Aaa spread is the difference between BAA- and AAA-rated corporate bond yields, $r^{BAA} - r^{AAA}$, which is a measure of market credit risk.

A set of firm characteristics that are related to stock return volatility is considered. First, we account for a firm's *market capitalization*, MV, measured as the product of the closing price on the last trading day of a month and the number of shares outstanding. Moreover, it has been well-documented that an increase in the proportion of debt to equity leads to an increase in return volatility (see Merton (1974) and Schwert (1989), for example), and hence, we include *leverage* as the ratio of total liabilities to total assets. In addition, the *earnings-price ratio* (EPS/P) and *dividend-price ratio* (D/P) are shown to be closely associated with return volatility (see Mele (2007) and Christiansen et al. (2012), for example). Furthermore, Schwert (1989) provides evidence that there is a strong association between stock return volatility and stock trading activity. To capture this effect, we include monthly *trading value*, measured as the product of the closing price on the last trading day of a month and the total number of shares sold during the month.

2.3 Data

2.3.1 Customer-Supplier Relationships

According to the U.S. Statement of Financial Accounting Standards (SFAS) No.131, public enterprises are required, once each year, to report the customers that account for at least 10% of their total yearly sales. This information is contained in the Compustat Customer Segment files. For each supplier, the key items in each entry of the customer segment files are the customer's name and the total amount of annual sales from this supplier to this customer.

As major customers are self-reported and, in particular, names are manually entered, the matching of a reported customer's name with a standard identifier is not a straightforward matter. For example, the same company can be reported with different names (IBM vs. International Business Machines), acronyms are included in some instances and omitted in others, or the company's name can be outright misspelled. We adopt a very conservative approach—we only consider those customer-supplier relations (i.e., links) for which there is an exact match (case-insensitive) between the reported name and an entry, which can be the company name or the company's legal name, or ticker, in the Compustat datafile of names.

Following this procedure, in the period from June 1976 to October 2015, 7,459 unique firms with a total of 40,279 reported links are identified. That is, during this period, there are, in total, 7,459 unique firms that have ever appeared in the customer-supplier networks identified from the customer segment files. Among them, 5,078 firms reported their customers, 3,264 firms were reported as customers by other companies, and 883 firms took the role of both a supplier and a reported customer.

Public companies are required to report their major customers once every fiscal year. As fiscal years vary across businesses, they report in different months. When a link is reported, we consider it active for a one-year period.[12] Specifically, it is considered active for up to one year prior to the reporting date.[13]

2.3.2 Other Data

To calculate realized stock return volatilities, we collect daily closing stock prices from the Center for Research in Security Prices (CRSP) U.S. Stock database for the period from January 1977 to December 1982 and from the Compustat-North America database for the period from January 1983 to December 2015. As our customer-supplier links are identified from the customer segment files of the Compustat-North America database, it is natural to collect stock prices from the same source; however, because daily data prior to January 1983 are not available there, we collect daily prices from the CRSP U.S. Stock database for the earlier period. We focus our analysis on common stocks and American depositary receipts (ADR) only, which together account for approximately 90% of the daily closing stock price observations in the sample. For firms that have multiple issues, we use that with the highest average daily trading volume throughout the sample period as the representative issue.[14] Dividends are reinvested. We exclude the observations with a stock price that is less than $1 and those for which the number of outstanding shares differs from the previous trading day. If there is a suspension longer than 7 days, the daily price observation on the day immediately after the suspension is also excluded.

Monthly interest rates on Treasury constant maturities and corporate bonds are collected from the Federal Reserve Bank database. Daily returns on the Standard & Poor's 500 Composite Index and monthly data on closing stock price, number of shares outstanding and number of shares traded are from the CRSP. Quarterly data on a firm's total assets, total liabilities, earnings per share (excluding extraordinary items), and monthly dividends per share are collected from the Compustat-North America database. Daily value-weighted returns on industry portfolios are obtained from Kenneth French's website.[15]

Table 2.1 contains the summary statistics for our final sample. Our data have a panel structure covering the period from February 1977 to October 2015. After matching with data on the entire set of variables listed in Section 2.2 - 2.2.3, there are 2,738 unique suppliers with a total of 134,007 monthly observations. The panel is unbalanced: the number of monthly observations for each supplier varies between 1 and 316, with a median of 34.

[Insert Table 2.1 about here.]

2.4 Estimation Results

With various sets of control variables, Equation 2.2 is estimated by pooled ordinary least squares (OLS) with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence (Driscoll and Kraay (1998)). Except for four of the variables, which are the yield curve slope, Baa-Aaa spread, earnings-price ratio (EPS/P) and dividend-price ratio (D/P), all variables in Equation 2.2 are natural-logarithm transformed prior to estimation.[16] The results are presented in Table 2.2. Motivated by the existing theoretical and empirical literature, after controlling for the variables capturing the time-series properties of stock return volatility and a set of market and idiosyncratic factors, customer volatility is a statistically significant determinant of a supplier's volatility—the estimated coefficient is significant at the 0.1% level in all specifications. In addition, according to Model (4), the estimated effect of customer volatility is approximately 10 times as large as trading value on supplier's volatility. Further, the effect of customer volatility is approximately one-third and one-fifth of the effect of RV^m in Model (4) and (5), respectively.[17] The predictive power of customer volatility is to be further examined in the out-of-sample analysis in Section 2.6.

[Insert Table 2.2 about here.]

2.5 Robustness

2.5.1 Industry Effects

In addition to originating from customer-supplier linkages, an alternative explanation for the presence of a customer effect in our framework is an industry effect. Averaging over customers' stock return volatilities, the argument holds, builds proxies for return volatility of the industry in which this firm operates. To address this concern, we introduce industry volatility as an additional control variable.

We obtain daily value-weighted stock returns on industry portfolios from Kenneth French's website.[18] These returns are constructed by assigning each AMEX, NYSE and NASDAQ stock to an industry portfolio according to its Standard Industrial Classification (SIC) code. For robustness, we consider various classifications, resulting in 12, 17, 30, 38 and 48 industry portfolios. For each classification scheme and each industry portfolio, first, we compute industry volatility as the sample standard deviation of daily returns in a month. Second, given a classification scheme, each firm in our dataset is matched to its industry portfolio according to its Compustat SIC code. Let $indRV_i$ denote the return volatility of the industry portfolio to which firm i is matched. Controlling for industry volatility, partial effects are estimated for the following model:

$$RV_{i,t} = \beta_0 + \beta_1 (G \cdot RV)_{i,t-1} + \beta_2 \; indRV_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t}, \tag{2.3}$$

where *ControlVariables* includes the entire set of variables listed in Section 2.2 - 2.2.3.

As presented in Table 2.3, the effects of industry volatility are positive and statistically significant. Compared to the results presented in Table 2.2, the estimated coefficient on customer volatility decreases very marginally (with respect to Model (5) in Table 2.2 which contains the entire set of control variables other than industry volatility), and it is statistically significant at the 0.1% level for all of the industry classification schemes considered.

[Insert Table 2.3 about here.]

2.5.2 Activeness of Customer-Supplier Linkages

As required by SFAS no.131, public companies report their major customers once every fiscal year. As fiscal years vary across businesses, they report in different months during the year. We consider a customer-supplier linkage active for up to a one-year period once a supplier company reports the name of and the yearly sales to the customer company. In this section, we compare three windows for the activeness of customer-supplier linkages and thus for the construction of sales-weighted G: (1) one year prior to the reporting date; (2) one year centered on the reporting date; and (3) one year after the reporting date.[19] For each of the three windows, Equation 2.2 is estimated by pooled OLS with the Driscoll-Kraay standard error. The estimated effects of customer volatility, as reported in Table 2.4, are quite stable across different choices of windows.

[Insert Table 2.4 about here.]

2.5.3 Estimation with the Hausman-Taylor Approach

When including an autoregressive term (RV^m), Equation 2.2 is a dynamic panel model, and a traditional fixed effects (de-meaned) regression would cause endogeneity by design; correlation between the lagged dependent variable and the unobserved individual fixed effect is non-zero. In other words, if one believes that an unobserved individual time-invariant effect (u_i) exists, it must be the case that $Cov(y_{t-1}, u_i) \neq 0$, which would certainly cause inconsistency. Therefore, in this section, we estimate Equation 2.2 by the Hausman-Taylor approach, which controls for the potential correlation between the individual fixed effect and the autoregressive term (Hausman and Taylor (1981)).[20]

As reported in Table 2.5, estimated coefficients on customer volatility under various specifications are generally larger than those obtained above using pooled OLS and are all statistically significant at the 0.1% level. In particular, the magnitude of the effect from customer volatility is still about 10 times as large as that from trading value on supplier's volatility in Model (4), (5) and (6).

[Insert Table 2.5 about here.]

2.5.4 Estimation with Instrumental Variables

There might be concerns that the variable of interest, customers' return volatility, is endogenous (i.e, not orthogonal to the error term). For example, one concern is that the regression equation is simultaneous in the sense that the transfer of volatility also goes from suppliers to customers. First, we argue that this simultaneity is unlikely because our investigation is undertaken in a predictive regression setting—supplier's volatility, the dependent variable, is one-period later in time than the customer volatility. However, regardless of the source of endogeneity, we explicitly address this concern by estimating Equation 2.2 by constructing and utilizing instrumental variables (IVs) in this section.

First, we construct $G_{i,t}^{new,k}$ to capture firm i's newly established customers in month t. Specifically, $G_{i,t}^{new,k}$ captures those customers of firm i in month t that were not firm i's customers k months ago, that is, firms that while they are firm i's customers in month t are not customers of firm i in month $t-k$. Next, RV_{t-k}, the $n \times 1$ vector containing each firm's return volatility in month $t-k$, is multiplied by $G_t^{new,k}$, to produce the IV. That is, the k-month-lagged return volatility of these "newly established customers", $(G_t^{new,k} \cdot RV_{t-k})_i$, is utilized as an IV. As these firms are not firm i's customers k months ago by construction, their k-month-lagged volatilities are *unlikely* to be correlated with the current disturbance term. Hence, we claim this IV to be exogenous. However, as return volatility is in general a persistent process, k-month-lagged volatilities of these newly established customers should be correlated with their own more recent volatilities; hence, this IV is correlated with the variable of interest.

The estimation results are presented in Table 2.6. Specifically, we use different combinations of the following IVs: $(G_{t-2}^{new,2} \cdot RV_{t-4})_i$, $(G_{t-2}^{new,6} \cdot RV_{t-8})_i$ and $(G_{t-2}^{new,12} \cdot RV_{t-14})_i$. They capture the corresponding lagged return volatilities of firm i's newly established customers that are not a customer of firm i, 2 months, 6 months and 12 months ago, respectively. The coefficients are estimated by the two-step efficient Generalized Method of Moments

(GMM) procedure, with the estimated asymptotic variance of the GMM estimator being heteroskedasticity and autocorrelation consistent (HAC). The asymptotic variance of the sample analogue of the orthogonality or moment conditions (specifying that all of the instruments in the equation are uncorrelated with the error term) is estimated using a Bartlett kernel with bandwidth $q(n) = 13$ (Newey and West (1987)). The inverse of the estimated asymptotic variance is then used as the weighting matrix in the second stage of the GMM estimation to obtain the efficient GMM estimator[21].

[Insert Table 2.6 about here.]

We report the Kleibergen-Paap Wald rk F statistic (Kleibergen and Paap (2006) and Kleibergen and Schaffer (2015)) as a weak-instruments test. With every combination of IVs, the F statistic is above 700—it well exceeds the "rule of thumb" requirement of Staiger and Stock (1997), which states that the F statistic should be greater than 10 for weak identification not to be considered a problem. We also report the p-value that is associated with Hansen's test of overidentifying restrictions (Hansen (1982)). The large p-values, ranging from approximately 0.33 to 0.56, indicate in our favor that we *fail to reject* the null that *all the regularity assumptions of the model (including the assumption that the IVs are orthogonal to the error term) are satisfied.* Most important, the estimated coefficients on customer volatility are comparable to the previous results, in terms of sign, magnitude and level of statistical significance. That is, our previous results are further confirmed by the estimation with IVs.

2.5.5 Stock Return Volatility and Randomized Linkages

To further demonstrate the robustness of our results, in this section, we construct $(G^R \cdot RV)$, where G^R is the randomized sales-weighted G: columns of sales-weighted G are shuffled randomly[22]. That is, rather than capturing firm i's customers, the ith row of G^R contains randomly selected firms which may or may not be firm i's customers. So the ith entry in $(G^R_t \cdot RV_t)$ is the weighted average of the return volatilities of randomly selected firms in month t.

With various sets of control variables, the following equation is estimated by pooled OLS with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence:

$$RV_{i,t} = \beta_0 + \beta_1 \left(G^R \cdot RV\right)_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t}, \tag{2.4}$$

Estimation results are reported in columns (2) and (4) of Table 2.7. For comparison, benchmark results from Table 2.2 are presented in columns (1) and (3), where customer volatility $(G \cdot RV)$ is constructed using sales-weighted G. Comparing to the estimated coefficients on the customer volatility, the coefficients on $(G^R \cdot RV)$ are smaller in magnitude, and more importantly, do not have statistical significance. That is, results in Table 2.7 indicate that, while customers' return volatility predicts that of suppliers, return volatilities of randomly selected firms are not significant determinants of a firms's volatility.

[Insert Table 2.7 about here.]

2.6 Out-of-Sample Analysis

2.6.1 Out-of-Sample Forecasting Tests

In this section, we test whether incorporating customer volatility improves volatility forecasts from an out-of-sample perspective. Let model 1 be the benchmark model; model 2, which nests model 1 and includes customer volatility as an additional forecasting variable, is the augmented model.

$$\text{Model 1: } RV_{i,t} = \beta_0 + X_{i,t-1}\boldsymbol{\beta}_1 + \epsilon_{i,t}, \tag{2.5}$$

$$\text{Model 2: } RV_{i,t} = \beta_0 + X_{i,t-1}\boldsymbol{\beta}_1 + \beta_2 \left(G \cdot RV\right)_{i,t-1} + \epsilon_{i,t}, \tag{2.6}$$

where $X_{i,t-1}$ is a row vector of forecasting variables of firm i in month $t-1$, which does not include customer volatility, $(G \cdot RV)_{i,t-1}$; $\boldsymbol{\beta}_1$ is a column vector of parameters.

Mean squared prediction error (MSPE) has been one of the most commonly used statistics for comparing forecasts of an augmented model to the nested benchmark model (see, for example, Lettau and Ludvigson (2001), Stock and Watson (2002), Stock and Watson (2003), Stock and Watson (2004) and Orphanides and van Norden (2005)). We apply the test for equal MSPE proposed by Clark and West (2007). Specifically, Clark and West (2007)'s approach centers on the idea that, under the null that the benchmark model generates the data, the augmented model introduces noise into the forecasting process by estimating parameters with population values of zero; hence, the MSPE of the augmented model is expected to be larger than that of the benchmark model. The MSPE of the augmented model should therefore be adjusted to account for this noise. In our context, the sample MSPE for model 1 and model 2 are, respectively,

$$\hat{\sigma}_1^2 = P^{-1} \sum_i \sum_t (RV_{i,t+1} - \hat{RV}_{1,i,t+1})^2, \tag{2.7}$$

$$\hat{\sigma}_2^2 = P^{-1} \sum_i \sum_t (RV_{i,t+1} - \hat{RV}_{2,i,t+1})^2, \tag{2.8}$$

where $\hat{RV}_{1,i,t+1}$ and $\hat{RV}_{2,i,t+1}$ denote one-step-ahead predictions of $RV_{i,t+1}$ from model 1 and 2, respectively; P denotes the number of out-of-sample predictions used in computing these averages. The adjustment, which is subtracted from $\hat{\sigma}_2^2$ to account for the additional noise associated with the augmented model's forecasts, is

$$adj. = P^{-1} \sum_i \sum_t (\hat{RV}_{1,i,t+1} - \hat{RV}_{2,i,t+1})^2. \tag{2.9}$$

Clark and West (2007) propose testing the null hypothesis of equal MSPE by examining the "MSPE-adjusted" statistic:

$$MSPE\text{-}adjusted \equiv \hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.). \tag{2.10}$$

This is a one-sided test with the alternative hypothesis that model 1 has a greater MSPE than model 2. Clark and West (2007) argue that, although the MSPE-adjusted statistic is not asymptotically normal, standard normal critical values result in actual sizes close to, but slightly smaller than, nominal size for sufficiently large samples.

For the out-of-sample analysis, the entire sample is divided into two parts: the observations in the first T_1 time periods are for estimating the parameters in the forecasting models (estimation sample), and the observations in the final T_2 time periods are used to estimate the MSPE associated with each model (forecasting sample). We use 24-, 36-, 48- and 60-month rolling estimation windows to estimate the parameters in the forecasting models.

We select seven benchmark models (i.e., model 1) for forecast evaluation. First, following Paye (2012), an AR(6) specification is used:

$$RV_{i,t} = \beta_0 + \sum_{k=1}^{6} \beta_k RV_{i,t-k} + \epsilon_{i,t}. \tag{2.11}$$

The next four benchmark models each contains one set of the following forecasting variables: heterogeneous autoregressive terms (RV^m, RV^q, RV^{hy}, and RV^y); market factors ($RV_{S\&P}$, yield curve slope and Baa-Aaa spread); firm characteristics (MV, leverage, EPS/P, D/P and trading value); and industry volatility under the 30-industry classification scheme. The sixth benchmark model contains heterogeneous autoregressive terms, market factors and industry volatility that are as specified above. The last benchmark model includes the entire set of variables introduced in Section 2.2 - 2.2.3 and industry volatility as forecasting variables.

In Table 2.8, the Clark and West MSPE-adjusted statistic and corresponding p-value are reported for each benchmark and estimation scheme combination. For the first six benchmark models, the null of equal MSPE is rejected for all estimation schemes used, most of which at the 0.1% significance level. Even with the most aggressive choice, namely, using a benchmark model containing the entire set of variables introduced in Section 2.2 - 2.2.3 and industry volatility as forecasting variables (*All-But-No-CRV*), the null of equal MSPE is rejected at the 5% level when using the 60-month rolling estimation window. These results provide consistent evidence that incorporating customer volatility improves forecasts of supplier volatility.

[Insert Table 2.8 about here.]

2.6.2 Application: Density Forecasts

Density forecasting plays an important role in financial risk management, for example, in measuring and monitoring asset or portfolio Value-at-Risk. Coupled with the assumption that a firm's daily stock returns are normally distributed, the customer-channel-based volatility forecast can be used to forecast stock return density. Following Andersen, Bollerslev, Diebold, and Labys (2003), we assess our density forecasts using the methods of Diebold, Gunther, and Tay (1998). Suppose that the daily stock returns, r_τ, in month t follows conditional density, $f(r_\tau \mid \mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ denotes the full information set available in month $t-1$. If the forecasted density, $f_{t|t-1}(r_\tau)$, is equal to the true density, $f(r_\tau \mid \mathcal{F}_{t-1})$, the sequence of probability integral transforms of daily returns with respect to $f_{t|t-1}(\cdot)$[23] should be uniformly distributed on $(0,1)$. Thus, the performance of the density forecasts and, thus, also the customer-channel-based volatility forecasts, can be investigated by checking whether the distributions of the probability integral transforms are $U(0,1)$.

In Table 2.9, we report six selected quantiles of the probability integral transforms of observed daily returns using the sequence of one-month-ahead predicted volatilities. Again, we use 24-, 36-, 48- and 60-month rolling windows to estimate the parameters in the forecasting models. For all schemes, the first estimation sample starts in December 1984. As reported, for each estimation scheme, the percentages are approximately consistent with the corresponding selected quantiles, which indicates that, when coupled with the assumption that a firm's daily stock returns are normally distributed, our customer-volatility model generally performs well at forecasting stock return density.

[Insert Table 2.9 about here.]

2.7 Stock Return Volatility and Analyst Coverage

In addition to demonstrating the predictive power of customer return volatility, we find the following question quite intriguing: when investors are more aware of the identity of a firm's principal customers, will we observe more pronounced volatility transferring through the customer channel? To gain insights into this question, we investigate the interaction between customer volatility and analyst coverage. Intuitively, the research activities and recommendations of financial analysts reveal a firm's information to the market. Ceteris paribus, a higher number of analysts covering a firm entails a higher level of information transparency for this firm; that is, the public has better access to information regarding its operations and financial conditions and is more likely to be aware of its major customer-supplier relationships.

Specifically, we estimate the following model:

$$\begin{aligned} RV_{i,t} = \beta_0 &+ \beta_1 \left(G \cdot RV\right)_{i,t-1} \\ &+ \beta_2 \left(G \cdot RV\right)_{i,t-1} \times N^{Analysts}_{i,t-1} + \beta_3 N^{Analysts}_{i,t-1} \\ &+ ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t}, \end{aligned} \tag{2.12}$$

where $N^{Analysts}_{i,t}$ is the number of analysts who issue recommendations (which can be buy, hold, or sell) for firm i in month t. We also use $Ave.\ N^{Analysts}$ as an alternative measure of analyst coverage, which is the monthly average number of analysts who issue recommendations for a firm over the sample period. The numbers of recommendations are collected from the Institutional Brokers' Estimate System (I/B/E/S) database. Given the availability of data, our sample covers the period from December 1993 to October 2015.

Equation 2.12 is estimated by pooled OLS with the Driscoll-Kraay standard error. The results are reported in Tables 2.10 and 2.11. In Table 2.10, MV and trading value are included to control for the fact that larger companies tend to have higher analyst coverage, and an industry dummy variable under the 12-industry classification scheme is included to capture the various industry properties. In Table 2.11, the entire set of variables listed in Section 2.2 - 2.2.3 and industry volatility under the 30-industry classification scheme are included as controls. Variables other than the industry dummy, yield curve slope, Baa-Aaa spread, EPS/P, D/P, $N^{Analysts}$ and $Ave.\ N^{Analysts}$ are log-transformed prior to estimation. For robustness, Equation 2.12 is also estimated by the Hausman-Taylor approach—the results are reported in columns (5) and (6) of Table 2.11.

[Insert Table 2.10 about here.]

[Insert Table 2.11 about here.]

In most cases, the estimated coefficient on the interaction term is positive and statistically significant at the 0.1% level. These results show that a higher number of analysts who follow a firm is generally associated with larger customer effect on suppliers, which indicates that the transfer of volatility from customers to suppliers is more pronounced when investors are more aware of the linkages. Our interpretation is that for a firm with higher analyst coverage, investors are likely to be more aware of its principal customers; as they anticipate the propagation of shocks through customer-supplier linkages, they respond more actively to news regarding firm's customers. That is, they incorporate such news into their investment and asset allocation decisions related to the supplier companies. This trend contributes to the stronger association between customer and supplier volatility. Our observation and interpretation shed light on one of the mechanisms of shocks propagation via customer-supplier linkages; and indicate that customer-supplier relationship is a channel for news to be incorporated into stock prices.

2.8 Conclusions

The majority of the existing literature has focused exclusively on aggregate equity market volatility, or pure time-series modeling and forecasting of volatility. In contrast, this paper investigates stock return volatility from the perspective of individual firms and examines one particular channel—the transfer of volatility along supply chains, in a predictive regression setting.

Existing literature documents that there is a strong association between stock return volatility and trading activity. Our analysis shows that the effect of customer volatility is approximately 10 times as large as trading value on supplier's volatility. Our findings are robust to controlling for variables capturing the time-series properties of stock return volatility and a set of idiosyncratic, industry, and market factors. Further, they are tested under various assumptions regarding the activeness of customer-supplier linkages and using different estimation methods, including estimations with instrumental variables using GMM. Moreover, using various benchmarks and rolling estimation windows, our out-of-sample tests produce consistent evidence of improvements in volatility forecasting when the customer channel is included.

In addition, we demonstrate that the transfer of volatility through the customer channel is more pronounced for firms with higher financial analyst coverage, after controlling for the size of firm, measured by market capitalization and trading value. This result is consistent with our expectation that with greater awareness of the identity of a firm's major customers, investors tend to respond more actively to news regarding these customers by incorporating such news into their investment decisions related to the supplier company. This trend contributes to the stronger association between customer and supplier volatility.

2.9 Notes

[8]Details regarding the data and the matching procedure that we use to identify the customer-supplier relations and to construct the sales-weighted G are provided in Section 2.3 - 2.3.1.

[9]The adjacency matrices that characterize customer-supplier networks are constructed as in Gençay et al. (2015).

[10]We also construct $RV_{i,t}$ without subtracting the sample average of daily returns—this version of stock return volatility produces quite similar estimation results to those obtained using the de-meaned version, for all specifications in this paper.

[11]Further information about this data source can be found in Section 2.3 - 2.3.1.

[12]We may have a different sales-weighted G in every month, as there are some firms reporting in different months, but each row of the sales-weighted G (i.e., each firm's customer linkages) is fixed for a 12-month period.

[13]In Section 2.5 - 2.5.2, we consider alternative window periods for the activeness of customer-supplier linkages, that is, for up to one year centered on the reporting date and one year after the reporting date.

[14]For each issue, average daily trading volume is measured as its total number of shares traded divided by its total number of trading days in the entire sample. Hence, by construction, each issue has an unchanged average daily trading volume throughout the entire period.

[15]Industry variables and data are explained in greater detail in Section 2.5 - 2.5.1.

[16]To preserve the sample size, the yield curve slope, Baa-Aaa spread, EPS/P and D/P are not log-transformed, as they have some non-positive observations.

[17]We also use total assets instead of MV to measure a firm's size and the total number of shares traded instead of trading value in a month to measure a stock's trading activity; the estimation results for customer volatility are quite similar with these different choices of control variables.

[18]These data and definitions are available online at Kenneth French's website: http://mba.tuck.dartmouth.edu /pages/faculty/ken.french/index.html.

[19]In the results presented previously, we used the first type of window "one year prior to the reporting date" for the activeness of customer relations.

[20]It should be noted that the Hausman-Taylor estimator does not remove the individual fixed effect. For the estimator to be consistent, we need to assume that there is no correlation between the individual fixed effect and any variables other than the dependent variable.

[21]See Hayashi (2000), Sections 3.5 and 6.6, for further details.

[22]Columns of sales-weighted G are shuffled randomly by the MATLAB fucntion "randperm" with the seed of randomization set to 100.

[23]In particular, this is the cumulative density function with respect to $f_{t|t-1}(\cdot)$ evaluated at the observed daily return, r_τ: i.e., $\int_{-\infty}^{r_\tau} f_{t|t-1}(u)du$.

2.10 Tables and Figure

Table 2.1: **Summary Statistics**

This table contains summary statistics for the main variables in our final sample (without log-transformation). The data have a panel structure covering the period from February 1977 to October 2015. There are 2,738 unique suppliers with a total of 134,007 monthly observations. The panel is unbalanced: the number of monthly observations for each supplier varies between 1 and 316, with a median of 34. RV, realized return volatility, is the sample standard deviation of daily returns over one month; $(G \cdot RV)$, customer volatility, is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). RV^q, RV^{hy}, and RV^y are realized return volatilities over one quarter, one-half year and one year, respectively. MV, market capitalization, is the product of the closing price on the last trading day of a month and the number of shares outstanding. *Leverage* is the ratio of total liabilities to total assets. EPS/P and D/P are the earnings-price ratio and dividend-price ratio, respectively. *Trading value* is the product of the closing price on the last trading day of a month and the total number of shares sold during the month. MV and trading value are reported in millions and thousands of dollars, respectively. For comparison, we also report statistics for the entire sample (the "universe"), spanning the same period as ours, from which data on each variable are collected. These statistics are presented in square brackets underneath the corresponding variables.

	Mean	**Std. Dev.**	**Min.**	**Max.**	**N**
$(G \cdot RV)$	0.020	0.016	0	1.891	134,007
RV	0.034	0.025	0	2.437	134,007
	[0.037]	[0.047]	[0]	[27.997]	[2,634,700]
RV^q	0.035	0.023	0.003	1.623	134,007
	[0.036]	[0.035]	[0]	[16.213]	[2,542,397]
RV^{hy}	0.036	0.023	0.005	1.623	134,007
	[0.035]	[0.031]	[0]	[11.182]	[2,407,074]
RV^y	0.035	0.020	0.001	0.773	134,007
	[0.034]	[0.023]	[0]	[1.829]	[2,198,866]
MV ($ million)	2,863	12,837	0.322	501,512	134,007
	[2,153]	[11,816]	[0.005]	[750,709]	[1,993,997]
Leverage	0.455	0.347	0.010	22	134,007
	[1.972]	[65.663]	[-4,927]	[25,968]	[3,060,729]
EPS/P	-0.030	3.074	-647.498	6.234	134,007
	[-7.943]	[3,551.363]	[-2,762,672.750]	[258,979.359]	[3,026,240]
D/P	0.001	0.010	0	1.846	134,007
	[0.009]	[13.230]	[0]	[27,375]	[4,307,487]
Trading Value ($ 1,000)	546,592	2,174,322	3.487	113,449,435	134,007
	[312,754]	[1,945,809]	[0.021]	[344,843,943]	[1,993,838]

Table 2.2: **Stock Return Volatility and Customer-Supplier Linkages**

Partial effects are estimated with various sets of control variables for the following model:

$$RV_{i,t} = \beta_0 + \beta_1 \left(G \cdot RV\right)_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, and customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). An autoregressive term (RV^m) and the heterogeneous autoregressive terms—volatilities over different time horizons—quarter (RV^q), half year (RV^{hy}) and year (RV^y), are included. Market factors include the market volatility ($RV_{S\&P}$), yield curve slope ($r^{10} - r^2$) and Baa-Aaa spread ($r^{BAA} - r^{AAA}$). Firm characteristics include market capitalization (MV), leverage (total liabilities/total assets), earnings-price ratio (EPS/P), dividend-price ratio (D/P), and trading value. All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed. The data have a panel structure covering the period from February 1977 to October 2015. There are 2,738 unique suppliers with a total of 134,007 monthly observations according to model (5). The models are estimated by pooled OLS with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence (number in parentheses).

	(1)	(2)	(3)	(4)	(5)
$(G \cdot RV)$	0.3368***	0.0481***	0.0266***	0.0618***	0.0345***
	(0.0311)	(0.0128)	(0.0056)	(0.0126)	(0.0054)
RV^m		0.2119***	0.1985***	0.1922***	0.1726***
		(0.0168)	(0.0144)	(0.0170)	(0.0137)
RV^q		0.2581***	0.2509***	0.2531***	0.2443***
		(0.0252)	(0.0230)	(0.0226)	(0.0205)
RV^{hy}		0.3026***	0.3098***	0.2610***	0.2688***
		(0.0372)	(0.0331)	(0.0346)	(0.0303)
RV^y		0.0886***	0.0933***	0.0806***	0.0853***
		(0.0232)	(0.0206)	(0.0223)	(0.0202)
$RV_{S\&P}$			0.0702*		0.0875**
			(0.0307)		(0.0309)
YieldCurve Slope			−0.0212*		−0.0178
			(0.0104)		(0.0101)
BaaAaa Spread			−0.0390		−0.0349
			(0.0270)		(0.0236)
MV				−0.0336***	−0.0391***
				(0.0050)	(0.0052)
Leverage				−0.0022	−0.0037
				(0.0035)	(0.0032)
EPS/P				−0.0005	−0.0005
				(0.0003)	(0.0003)
D/P				−0.2604	−0.2827
				(0.1584)	(0.1595)
Trading Value				0.0061	0.0097**
				(0.0032)	(0.0035)
$\bar{R}^2$	0.10	0.60	0.60	0.58	0.59
N	181,004	160,932	160,932	134,007	134,007

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 2.3: **Controlling for Industry Effects**

Controlling for industry stock return volatility, partial effects are estimated for the following model:

$$RV_{i,t} = \beta_0 + \beta_1 \left(G \cdot RV\right)_{i,t-1} + \beta_2 \; indRV_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, and customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). Industry volatility ($indRV$) is the return volatility of the industry portfolio that the firm is matched to, which is based on one of the five classification schemes (12, 17, 30, 38, and 48 industry portfolios). Heterogeneous autoregressive terms include RV^m RV^q, RV^{hy}, and RV^y. Market factors include $RV_{S\&P}$, the yield curve slope and the Baa-Aaa spread. Firm characteristics include MV, leverage, EPS/P, D/P, and trading value. All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed. The data have a panel structure covering the period from December 1984 to October 2015, given the availability of SIC codes that are used to match a firm to its industry portfolio. The models are estimated by pooled OLS with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence (number in parentheses).

	(1)	(2)	(3)	(4)	(5)
$(G \cdot RV)$	0.0285∗∗∗	0.0296∗∗∗	0.0307∗∗∗	0.0299∗∗∗	0.0277∗∗∗
	(0.0047)	(0.0046)	(0.0047)	(0.0047)	(0.0045)
indRV					
12-Industry	0.0740∗∗∗				
	(0.0123)				
17-Industry		0.0555∗∗∗			
		(0.0123)			
30-Industry			0.0454∗∗∗		
			(0.0120)		
38-Industry				0.0508∗∗∗	
				(0.0120)	
48-Industry					0.0636∗∗∗
					(0.0110)
Other Control Variables					
Hetero. AR	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Market Factors	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Firm Charact.	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Robustness					
Heteroskedasticity	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Serial Correlation	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Cross-Sectional Dep.	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
$\bar{R}^2$	0.55	0.55	0.55	0.55	0.55
N	102,823	102,823	102,823	102,823	102,823

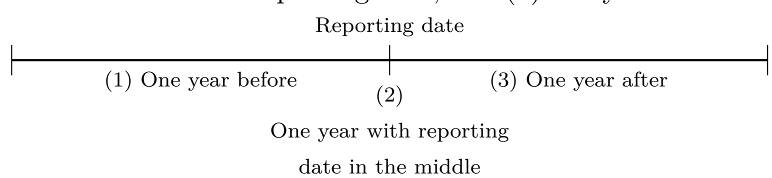
$*$ $p < 0.05$; $**$ $p < 0.01$; $***$ $p < 0.001$

Table 2.4: **Activeness of Customer-Supplier Linkages**

Partial effects are estimated for the following model:

$$RV_{i,t} = \beta_0 + \beta_1 (G \cdot RV)_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, and customer volatility ($G \cdot RV$) is constructed using sales-weighted G. A customer-supplier linkage is considered active for up to (1) one year prior to the reporting date; (2) one year centered on the reporting date; and (3) one year after the reporting date.



The entire set of control variables listed in Section 2.2 - 2.2.3 and industry volatility under the 30-industry classification scheme are included. All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed. The data have a panel structure covering the period from 1984 to 2015 (given the availability of SIC codes that are used to match a firm to its industry portfolio). The models are estimated by pooled OLS with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence (number in parentheses).

		(1)	(2)	(3)
$(G \cdot RV)$		0.0307∗∗∗	0.0308∗∗∗	0.0310∗∗∗
		(0.0047)	(0.0046)	(0.0045)
Control Variables				
	Hetero. AR	*Yes*	*Yes*	*Yes*
	Market Factors	*Yes*	*Yes*	*Yes*
	Firm Charact.	*Yes*	*Yes*	*Yes*
	Industry Vol.	*Yes*	*Yes*	*Yes*
Robustness				
	Heteroskedasticity	*Yes*	*Yes*	*Yes*
	Serial Correlation	*Yes*	*Yes*	*Yes*
	Cross-Sectional Dep.	*Yes*	*Yes*	*Yes*
$\bar{R}^2$		0.55	0.55	0.53
N		102,823	104,548	102,904

∗ $p < 0.05$; ∗∗ $p < 0.01$; ∗∗∗ $p < 0.001$

Table 2.5: **Estimation with the Hausman-Taylor Approach**

Partial effects are estimated with various sets of control variables for the following model:

$$RV_{i,t} = \beta_0 + \beta_1 (G \cdot RV)_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, and customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). An autoregressive term (RV^m) and the heterogeneous autoregressive terms—volatilities over different time horizons—quarter (RV^q), half year (RV^{hy}) and year (RV^y), are included. Market factors include the market volatility ($RV_{S\&P}$), yield curve slope ($r^{10} - r^2$) and Baa-Aaa spread ($r^{BAA} - r^{AAA}$). Firm characteristics include market capitalization (MV), leverage (total liabilities/total assets), earnings-price ratio (EPS/P), dividend-price ratio (D/P), and trading value. *indRV* is the return volatility of the industry portfolio that the firm is matched to, which is based on the 30-industry classification scheme. All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed. The data have a panel structure covering the period from 1984 to 2015. The models are estimated by the Hausman-Taylor approach which controls for the potential correlation between the individual fixed effect and the autoregressive term. The numbers in parentheses are standard errors.

	(1)	(2)	(3)	(4)	(5)	(6)
$(G \cdot RV)$	0.1519***	0.0978***	0.0512***	0.1041***	0.0557***	0.0458***
	(0.0025)	(0.0025)	(0.0028)	(0.0027)	(0.0030)	(0.0030)
RV^m		0.1713***	0.1500***	0.1615***	0.1339***	0.1280***
		(0.0038)	(0.0038)	(0.0043)	(0.0044)	(0.0044)
RV^q		0.2275***	0.2122***	0.2243***	0.2092***	0.2060***
		(0.0072)	(0.0071)	(0.0079)	(0.0078)	(0.0078)
RV^{hy}		0.2187***	0.2135***	0.1986***	0.1966***	0.1901***
		(0.0077)	(0.0077)	(0.0086)	(0.0085)	(0.0085)
RV^y		0.0521***	0.0597***	0.0477***	0.0535***	0.0497***
		(0.0048)	(0.0048)	(0.0053)	(0.0053)	(0.0053)
$RV_{S\&P}$			0.1146***		0.1223***	0.0538***
			(0.0035)		(0.0037)	(0.0054)
YieldCurve Slope			−0.0197***		−0.0214***	−0.0211***
			(0.0015)		(0.0016)	(0.0016)
BaaAaa Spread			0.0062		−0.0097**	−0.0076*
			(0.0035)		(0.0036)	(0.0036)
MV				−0.0320***	−0.0404***	−0.0455***
				(0.0034)	(0.0033)	(0.0033)
Leverage				0.0019	0.0022	0.0046
				(0.0030)	(0.0031)	(0.0031)
EPS/P				−0.0435***	−0.0418***	−0.0400***
				(0.0044)	(0.0044)	(0.0044)
D/P				0.0180	−0.0017	−0.0021
				(0.1156)	(0.1148)	(0.1146)
Trading Value				−0.0110***	−0.0052**	−0.0042*
				(0.0019)	(0.0019)	(0.0019)
indRV						0.0972***
						(0.0056)
N	142,958	127,084	127,084	102,823	102,823	102,823

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 2.6: **Estimation with Instrumental Variables**

Partial effects are estimated for the following model:

$$RV_{i,t} = \beta_0 + \beta_1 \left(G \cdot RV\right)_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, and customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). *ControlVariables* includes the entire set of variables listed in Section 2.2 - 2.2.3. We use different combinations of the following IVs: $(G_{t-2}^{new,2} \cdot RV_{t-4})_i$, $(G_{t-2}^{new,6} \cdot RV_{t-8})_i$ and $(G_{t-2}^{new,12} \cdot RV_{t-14})_i$. They capture the corresponding lagged return volatilities of firm i's newly established customers that are not a customer of firm i, 2 months, 6 months and 12 months ago, respectively. All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed.
Column (1) is estimated by pooled OLS with the Driscoll-Kraay standard error (in parentheses). Columns (2) - (4) are estimated using the two-step efficient GMM procedure, with HAC estimated asymptotic variance. The asymptotic variance of the sample analogue of the orthogonality conditions is estimated using a Bartlett kernel with bandwidth $q(n) = 13$. The inverse of the estimated asymptotic variance is then used as the weighting matrix in the second stage of the GMM estimation to obtain an efficient GMM estimator. The numbers in parentheses are HAC standard errors. We report the Kleibergen-Paap Wald rk F statistic as a weak identification test, with the null that the IVs and customer volatility are weakly correlated. We also report the p-value that is associated with Hansen's test of overidentifying restrictions, with the null that all of the regularity assumptions of the model (including the assumption that the IVs are orthogonal to the error term) are satisfied. The sample covers the period from 1977 to 2015.

	(1)	(2)	(3)	(4)
$(G \cdot RV)_{i,t-1}$	0.0345***	0.0534**	0.0518**	0.0497**
	(0.0054)	(0.0171)	(0.0176)	(0.0171)
Control Variables				
Hetero. AR	*Yes*	*Yes*	*Yes*	*Yes*
Market Factors	*Yes*	*Yes*	*Yes*	*Yes*
Firm Charact.	*Yes*	*Yes*	*Yes*	*Yes*
Instrumental Variables				
$(G_{t-2}^{new,2} \cdot RV_{t-4})$		*Yes*	*Yes*	*Yes*
$(G_{t-2}^{new,6} \cdot RV_{t-8})$		*Yes*		*Yes*
$(G_{t-2}^{new,12} \cdot RV_{t-14})$			*Yes*	*Yes*
Weak Identification Test, F statistic		719.714	723.429	733.803
Overidentifying Restrictions Test, p-value		0.3906	0.3266	0.5564
$\bar{R}^2$	0.59	0.61	0.61	0.61
N	134,007	7,238	7,155	7,151

$*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$

Table 2.7: **Stock Return Volatility and Randomized Linkages**

Partial effects are estimated with various sets of control variables for the following model:

$$RV_{i,t} = \beta_0 + \beta_1 (G \cdot RV)_{i,t-1} + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month. In columns (1) and (3), customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). In columns (2) and (4), ($G \cdot RV$) is constructed using randomized sales-weighted G: columns of sales-weighted G are shuffled randomly. An autoregressive term (RV^m) and the heterogeneous autoregressive terms—volatilities over different time horizons—quarter (RV^q), half year (RV^{hy}) and year (RV^y), are included. Market factors include the market volatility ($RV_{S\&P}$), yield curve slope ($r^{10} - r^2$) and Baa-Aaa spread ($r^{BAA} - r^{AAA}$). Firm characteristics include market capitalization (MV), leverage (total liabilities/total assets), earnings-price ratio (EPS/P), dividend-price ratio (D/P), and trading value. All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed. The data have a panel structure covering the period from February 1977 to October 2015. The models are estimated by pooled OLS with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence (number in parentheses). Columns (1) and (3) contain the benchmark results as in Table 2.2.

	(1)	(2)	(3)	(4)
$(G \cdot RV)$	0.0266***	0.0062	0.0345***	0.0040
	(0.0056)	(0.0035)	(0.0054)	(0.0033)
RV^m	0.1985***	0.1942***	0.1726***	0.1747***
	(0.0144)	(0.0152)	(0.0137)	(0.0149)
RV^q	0.2509***	0.2548***	0.2443***	0.2428***
	(0.0230)	(0.0251)	(0.0205)	(0.0231)
RV^{hy}	0.3098***	0.2990***	0.2688***	0.2636***
	(0.0331)	(0.0358)	(0.0303)	(0.0324)
RV^y	0.0933***	0.1044***	0.0853***	0.1004***
	(0.0206)	(0.0208)	(0.0202)	(0.0202)
$RV_{S\&P}$	0.0702*	0.0926**	0.0875**	0.1120***
	(0.0307)	(0.0324)	(0.0309)	(0.0327)
YieldCurve Slope	−0.0212*	−0.0265*	−0.0178	−0.0246*
	(0.0104)	(0.0109)	(0.0101)	(0.0108)
BaaAaa Spread	−0.0390	−0.0410	−0.0349	−0.0368
	(0.0270)	(0.0294)	(0.0236)	(0.0259)
MV			−0.0391***	0.0024
			(0.0052)	(0.0039)
Leverage			−0.0037	−0.0351***
			(0.0032)	(0.0054)
EPS/P			−0.0005	−0.0003
			(0.0003)	(0.0002)
D/P			−0.2827	−0.1078
			(0.1595)	(0.2896)
Trading Value			0.0097**	0.0093*
			(0.0035)	(0.0039)
$\bar{R}^2$	0.60	0.60	0.59	0.58
N	160,932	94,291	134,007	77,343

$*$ $p < 0.05$; $**$ $p < 0.01$; $***$ $p < 0.001$

Table 2.8: **Out-of-Sample Forecasting Tests**

This table reports the Clark and West MSPE-adjusted statistic and corresponding p-value (in parentheses) for testing the following hypothesis: $H_0 : MSPE_1 - MSPE_2 = 0; H_1 : MSPE_1 - MSPE_2 > 0$. $MSPE_1$ and $MSPE_2$ are the MSPE from model 1 (benchmark model) and 2 (augmented model), respectively:

$$\text{Model 1: } RV_{i,t} = \beta_0 + X_{i,t-1}\boldsymbol{\beta_1} + \epsilon_{i,t},$$
$$\text{Model 2: } RV_{i,t} = \beta_0 + X_{i,t-1}\boldsymbol{\beta_1} + \beta_2 \left(G \cdot RV\right)_{i,t-1} + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date), and X is a row vector of forecasting variables that does not include customer volatility. We select seven benchmarks: AR(6); heterogeneous autoregressive terms (RV^m, RV^q, RV^{hy}, and RV^y); market factors ($RV_{S\&P}$, yield curve slope and Baa-Aaa spread); firm characteristics (MV, leverage, EPS/P, D/P and trading value); industry volatility under the 30-industry classification scheme; a combination of heterogeneous AR, market factors and industry volatility; and, finally, a model includes the entire set of variables introduced in Section 2.2 - 2.2.3 and industry volatility as forecasting variables (*All-But-No-CRV*). All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed; 24-, 36-, 48- and 60-month rolling windows are used to estimate parameters in the forecasting models. For the first four benchmarks, the first estimation sample starts in February 1977; for the last three benchmarks that contain industry volatility, the first estimation sample starts in December 1984. The whole sample period is from February 1977 to October 2015.

	Rolling Estimation Window			
Benchmark Model	24-Month	36-Month	48-Month	60-Month
AR(6)	0.0169***	0.0138***	0.0142***	0.0090***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Hetero. AR	0.0161***	0.0136***	0.0137***	0.0136***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Market Factors	0.0023***	0.0022***	0.0038***	0.0060***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Firm Charact.	0.0844*	0.0365*	0.0582***	0.0560***
	(0.0428)	(0.0450)	(0.0000)	(0.0001)
Industry Vol.	0.0060***	0.0072***	0.0096***	0.0107***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
HAR-Market-Industry	0.0031***	0.0017**	0.0016***	0.0016***
	(0.0009)	(0.0018)	(0.0001)	(0.0000)
All-But-No-CRV	0.2184	−0.0278	0.0063	0.0062*
	(0.1286)	(0.3787)	(0.1828)	(0.0114)

One-sided test: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 2.9: **Distributions of Probability Integral Transforms**

This table reports the selected quantiles of the probability integral transform of returns with respect to the density forecasts from the following customer-volatility model:

$$RV_{i,t} = \beta_0 + \beta_1 \left(G \cdot RV\right)_{i,t-1} + X_{i,t-1}\Gamma + \epsilon_{i,t},$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date), and X is a row vector containing the entire set of variables listed in Section 2.2 - 2.2.3 and industry volatility under the 30-industry classification scheme. All variables, except for the yield curve slope, Baa-Aaa spread, EPS/P and D/P, are log-transformed; 24-, 36-, 48- and 60-month rolling windows are used to estimate parameters in the forecasting models. For all schemes, the first estimation sample starts in December 1984. The whole sample period is from December 1984 to October 2015.

	Rolling Estimation Window			
Quantile	24 Months	36 Months	48 Months	60 Months
5%	0.0256	0.0382	0.0437	0.0465
10%	0.0977	0.1089	0.1136	0.1158
25%	0.2914	0.2887	0.2888	0.2878
75%	0.7149	0.7184	0.7206	0.7226
90%	0.9156	0.9046	0.9008	0.8989
95%	0.9823	0.9717	0.9669	0.9646

Table 2.10: **Stock Return Volatility and Analyst Coverage I**

Partial effects are estimated for the following model:

$$\begin{aligned} RV_{i,t} &= \beta_0 + \beta_1 (G \cdot RV)_{i,t-1} \\ &+ \beta_2 (G \cdot RV)_{i,t-1} \times N^{Analysts}_{i,t-1} + \beta_3 N^{Analysts}_{i,t-1} \\ &+ ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t}, \end{aligned}$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, and customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). $N^{Analysts}$ is the number of analysts who issue recommendations (buy, hold, or sell) for a firm in a month (from the I/B/E/S database). We also use *Ave.* $N^{Analysts}$ as an alternative measure of analyst coverage, which is the monthly average number of analysts who issue recommendations for a firm over the sample period. MV and trading value are included to control for the fact that larger companies tend to have higher analyst coverage. An industry dummy variable under the 12-industry classification scheme is included to capture the various industry properties. All variables, except for the dummy, $N^{Analysts}$ and *Ave.* $N^{Analysts}$, are log-transformed. The sample covers the period from December 1993 to October 2015. The models are estimated by pooled OLS with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence (number in parentheses).

	(1)	(2)	(3)	(4)	(5)	(6)
$(G \cdot RV)$	0.3043***	0.2401***	0.2432***	0.2948***	0.2292***	0.2353***
	(0.0325)	(0.0291)	(0.0296)	(0.0325)	(0.0295)	(0.0294)
$(G \cdot RV) \times N^{Analysts}$	0.0044**	0.0056***	0.0044**			
	(0.0014)	(0.0014)	(0.0015)			
$N^{Analysts}$	−0.0005	0.0291***	0.0184**			
	(0.0060)	(0.0056)	(0.0061)			
$(G \cdot RV) \times$ *Ave.* $N^{Analysts}$				0.0065***	0.0064***	0.0052***
				(0.0013)	(0.0011)	(0.0012)
Ave. $N^{Analysts}$				0.0068	0.0390***	0.0282***
				(0.0055)	(0.0045)	(0.0050)
Control Variables						
MV and Trading Value		*Yes*	*Yes*		*Yes*	*Yes*
Industry Dummy			*Yes*			*Yes*
$\bar{R}^2$	0.18	0.33	0.30	0.17	0.34	0.30
N	104,577	94,106	76,564	104,577	94,106	76,564

$*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$

Table 2.11: **Stock Return Volatility and Analyst Coverage II**

Partial effects are estimated for the following model:

$$\begin{aligned} RV_{i,t} &= \beta_0 + \beta_1 \left(G \cdot RV\right)_{i,t-1} \\ &\quad + \beta_2 \left(G \cdot RV\right)_{i,t-1} \times N^{Analysts}_{i,t-1} + \beta_3 N^{Analysts}_{i,t-1} \\ &\quad + ControlVariables_{i,t-1}\Gamma + \epsilon_{i,t}, \end{aligned}$$

where realized return volatility (RV) is the sample standard deviation of daily returns over one month, and customer volatility ($G \cdot RV$) is constructed using sales-weighted G (a customer-supplier linkage is considered active for up to one year prior to the reporting date). $N^{Analysts}$ is the number of analysts who issue recommendations (buy, hold, or sell) for a firm in a month (from the I/B/E/S database). We also use $Ave.\ N^{Analysts}$ as an alternative measure of analyst coverage, which is the monthly average number of analysts who issue recommendations for a firm over the sample period. Heterogeneous autoregressive terms include RV^m, RV^q, RV^{hy}, and RV^y. Market factors include $RV_{S\&P}$, yield curve slope and Baa-Aaa spread. Firm characteristics include MV, leverage, EPS/P, D/P and trading value. $indRV$ is the industry volatility under the 30-industry classification scheme. Variables other than the yield curve slope, Baa-Aaa spread, EPS/P, D/P, $N^{Analysts}$ and $Ave.\ N^{Analysts}$ are log-transformed. The sample covers the period from December 1993 to October 2015. Columns (1) to (4) are estimated by pooled OLS with the Driscoll-Kraay standard error that is robust to heteroskedasticity and cross-sectional and temporal dependence; columns (5) and (6) are estimated by the Hausman-Taylor approach which controls for the potential correlation between the individual fixed effect and the autoregressive term. The numbers in parentheses are standard errors.

	(1)	(2)	(3)	(4)	(5)	(6)
$(G \cdot RV)$	0.1308***	0.0225***	0.1206***	0.0200**	0.0349**	0.0299***
	(0.0164)	(0.0063)	(0.0165)	(0.0063)	(0.0049)	(0.0052)
$(G \cdot RV) \times N^{Analysts}$	0.0040**	0.0007			0.0011*	
	(0.0014)	(0.0007)			(0.0004)	
$N^{Analysts}$	0.0221***	0.0044			0.0030	
	(0.0055)	(0.0027)			(0.0019)	
$(G \cdot RV) \times Ave.\ N^{Analysts}$			0.0050***	0.0010		0.0018***
			(0.0012)	(0.0006)		(0.0005)
$Ave.\ N^{Analysts}$			0.0319***	0.0067**		0.0095***
			(0.0047)	(0.0025)		(0.0024)
Control Variables						
Hetero. AR		*Yes*		*Yes*	*Yes*	*Yes*
Market Factors	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Firm Charact.	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Industry Vol.	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
$\bar{R}^2$	0.37	0.54	0.38	0.54		
N	93,058	70,509	93,058	70,509	70,509	70,509

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

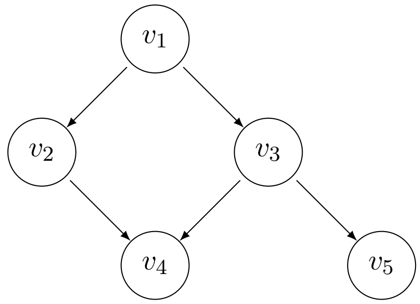


Figure 2.1: **A simple example of a customer-supplier network.**
In this figure, v_i, $i = 1, ..., 5$, denotes the firm; the arrow indicates the flow of output. For example, the arrow between v_1 and v_2 indicates that firm 1 (2) is the supplier (customer) of firm 2 (1).

Chapter 3

Application of Network Structures in Mutual Fund Performance Forecasting

Mutual fund companies play a major role in the financial markets, managing more than $15 trillion in assets, at year-end 2013, for nearly 98 million U.S. investors (Investment Company Institute, 2014). From 1980 to 2013, the percentage of American households owning mutual funds rose from 5.7% to 46.3%. In 2013, investors added $167 billion, as net cash inflows, to the mutual fund industry.

As is considered conventional wisdom and is documented in literature, mutual fund managers' stock selection skills are related to fund performance. Thus, persistent skill should lead to persistent performance: a good (bad) outcome today will be associated with good (bad) outcomes in the future. However, as the testing results usually depend on the choice of performance benchmark and investment horizon, empirically there is no consensus on whether performance is persistent or, if so, on what leads to persistence.

In this paper, instead of proving the existence of performance momentum[24] for the entire fund population, we argue that different funds have different abilities to maintain persistence of performance and that such abilities can be explained by how fund managers acquire and utilize market information in making investment decisions. That is, in addition to the conventional explanation for persistence in fund performance (manager's skill), we consider another source of heterogeneity among funds: the quality of information used by the fund manager. Specifically, managers' skills are constrained by the availability of market information: with unconstrained access to information, the good (bad) past performance of a fund is attributed to the good (bad) skill of a fund manager, resulting in a relatively persistent performance. With limited access to information, part of the outcome is attributed to luck, which is random, and thus, performance is less likely to persist overtime.

In our context, the term *information* is knowledge, in any form, that can potentially be transformed into profits. Examples include knowledge about which of two firms will win a lawsuit or whether an acquisition will take place (Dybvig and Ross, 1985); a New York Times article reporting on a new cancer drug from a drug company (Huberman and Regev, 2001); direct communications with publicly traded firms and brokerage firms through mutual fund companies' own investment banking, lending, and asset management divisions (Hendershott, Livdan, and Schürhoff, 2015); and consensus analyst recommendations (Kacperczyk and Seru, 2007). Due to time and budget constraints as well as the restrictions from the stated investment mandate, mutual fund companies collect information to different extents in terms of depth and broadness. Then, each fund manager applies his or her skill to utilize the information to make stock-holding decisions. Such a decision-making process can be considered a function of information and manager skill:

$$Decision_{i,t} = f_i(Information_{i,t}) = f(Information_{i,t}, Skill_i),$$

where $Information_{i,t}$ is a time-variant characteristic of a fund capturing how much information has been collected at that moment, and $Skill_i$ is the time-invariant ability of a fund to process the given information. This viewpoint regarding how a fund manager makes investment decisions is in line with the literature. For example, Dybvig and Ross (1985) express a portfolio manager's strategy as a function of information. Veldkamp (2006) argues that agents who purchase information in asset markets observe a component of the risky asset payoff, which guides their investments. Kacperczyk, Van Nieuwerburgh, and Veldkamp (2009) consider investment management business an information processing business. From the perspective of limited attention, Gupta-Mukherjee and Pareek (2014) argue that fund managers research assets in the choice set, and their performance depends on how efficiently they allocate attention among assets. Hendershott et al. (2015) states that superior information gathering and processing skills are important drivers of institutional trading.

To justify the hypothesis that the quality of information available to each fund is related to fund performance, we apply a new approach that incorporates network analysis into a mutual fund performance study. We construct a network of mutual funds based on the commonality of their stock holdings and use network features to characterize fund access to market information and to identify its relation with the strength of performance persistence. In particular, degree and clustering network coefficients are constructed as indices for mutual funds on the date they report their stock holdings. This design is inspired by the idea that common holdings of some stocks can reveal the commonality in market information to which a certain group of mutual funds has access. For example, there are research institutions that produce reports on companies, industries or markets. Such reports are purchased by various mutual fund companies, which may lead to common stock holdings and to the formation of a group of funds. With respect to information sources, different groups of funds represent different segments of the entire information set. If there is a fund with access to more different segments, that fund has more complete information than other funds do. Using stock holdings to reveal information possession can also be found in the body of literature on limited attention. For example, Barber and Odean (2008) argue that "investors do not buy all stocks that catch their attention; however, for the most part they only buy stocks that do so". Fang, Peress, and Zheng (2014) state, "an investor is unlikely to 'pull the trigger' on a stock trade unless he has paid some attention to the stock."

We argue that a network constructed based on commonality in stock holdings is informative in distinguishing funds in terms of their access to market information. Without any parametric model, degree and clustering network coefficients can be calculated and used as proxies to explain heterogeneity among funds in terms of their information accessibility and ability to generate persistent performance. For example, a mutual fund with a large clustering coefficient, by definition, has connections with other funds that share the same opinion and are thus connected themselves, indicating that a fund and the funds it connects to have access to (or rely on) a common source of information, such as general public media. Conversely, a fund with a small clustering coefficient is one that "listens to" different opinions—it connects to funds that do not have commonalities among themselves, indicating that the fund has access to different sources of information, such as valuable private information, and is therefore likely to possess more complete information about the market. The intuition is that a low level of clustering indicates that a fund has access to diversified market information sources. According to our explanation of mutual fund performance, that is, that fund manager decision making is constrained by information, the difference in information access will be reflected in mutual funds' performance persistence over time.

There are several studies that also examine the heterogeneity of information used by investors. Kacperczyk and Seru (2007) find that mutual fund managers respond to public and private information differently. Baker, Litov, Wachter, and Wurgler (2010) and Hendershott et al. (2015) show that mutual fund managers are informed (private information exists). Fang et al. (2014) and Yuan (2015) use market-wide, attention-grabbing events as proxies for media coverage, showing that there is heterogeneity among funds in terms of how they are affected by the change of market information.[25]

Using data on U.S. equity funds[26] from 2001 to 2014, we find that clustering and degree coefficients together explain mutual funds' performance persistence such that a mutual fund with more complete market information is more likely to possess momentum in performance. We also design an experiment to justify the validity of those two indices in representing a mutual fund's information-related characteristics.

This study extends the existing literature by providing a new explanation for mutual fund performance and incorporating stock holdings based network analysis into an empirical study. The remainder of this paper is organized as follows. Section 3.1 reviews the background and literature on mutual fund performance and related applications of network analysis. In Section 3.2, we explain how the two network-based indices are defined and constructed, and elaborate their implications. In Section 3.3, we present our empirical methodology and main results. Section 3.4 describes the design of the experiment and presents the results of the simulation. We conclude in Section 3.5.

3.1 Background and Literature Review

If mutual fund performance has momentum, a good (bad) outcome today will be associated with good (bad) outcomes in the future. Such a phenomenon has been documented and tested in literature, but the results are mixed and inconclusive. One important reason is that the choice of performance measure may differ; moreover, both past and future outcomes need to be considered in studying persistence, and these are not necessarily measured in the same way.

For example, Brown and Goetzmann (1995), Carhart (1997), Bollen and Busse (2004), Lückoff (2011), and Barroso and Santa-Clara (2015) use raw returns to capture past performance. A more popular choice is to apply some factor model-based risk-adjusted abnormal returns as the performance measurement. Examples include Grinblatt and Titman (1992), Brown and Goetzmann (1995), Elton, Gruber, and Blake (1996), Carhart (1997), Bollen and Busse (2004), and Barras, Scaillet, and Wermers (2010). Another study uses risk-adjusted abnormal returns as the performance measure but does not rely on a factor model (Grinblatt, Titman, and Wermers, 1995). Similarly, Barroso and Santa-Clara (2015) use the Sharpe ratio, which can be considered the same type of performance measure (ratio-based). Some studies apply mixed measures, using raw returns for past and risk-adjusted returns for future performance (Lesmond, Schill, and Zhou, 2004). There are also works that focus on designing new performance measures (Wermers, 2000; Cohen, Coval, and Pástor, 2005). Because judging the appropriateness of measures or proposing new measures of mutual fund performance is not our goal in this study, we select two common benchmarks, the Fama-French Three-Factor model (Fama and French, 1993) and CAPM (Lintner, 1965; Sharpe, 1964), to calculate abnormal returns for both past and future performance. Our results are robust to the choice between these two benchmarks.

Another issue that may affect performance measure is the length of the investment horizon considered. Elton et al. (1996), Bollen and Busse (2004), and Huij and Verbeek (2007) show that empirical results may change or even reverse when using different horizons. In this paper, we choose an investment horizon of one to five years when evaluating funds' performance.

In terms of the methodology for testing persistence, there are two main types: sorted portfolio methods (sorting funds into portfolios by past performance, then testing the difference in future performance among the sorted portfolios) and autoregressive models (regressing mutual funds' future performance on their past performance). Studies that use the former technique include Bollen and Busse (2004), Brown and Goetzmann (1995), Carhart (1997), Cohen et al. (2005), Elton et al. (1996), and Grinblatt et al. (1995), whereas Grinblatt and Titman (1992), Ter Horst and Verbeek (2000), and Bollen and Busse (2004) are examples that apply autoregressive models. We use both methods in our empirical study, and the results are robust to the choice of testing method. The procedure is detailed in Section 3.3.

One important innovation in our work is the application of mutual fund stock holdings data to explain the strength of performance persistence. Other works use the same data for different purposes. Grinblatt et al. (1995) study the dynamics of the weights of stocks in a mutual fund's portfolio and apply it to characterize the investment style of that fund. Cohen et al. (2005) argue that fund managers using similar techniques are likely to make similar decisions, and therefore, one can evaluate a fund manager's skill by the extent to which his investment decisions (stock holdings) resemble those of other successful managers. Pareek (2012) uses stock holdings data to study the pattern of stock returns, arguing that a stock held by a group of closely connected funds will respond to market-wide shocks more rapidly than other stocks. AntóN and Polk (2014) study the effect of common ownership on the correlations of stock returns. To the best of our knowledge, our work is the first to use mutual fund stock holdings data to measure funds' access to market information and to explain the persistence of mutual fund performance.

3.2 Network Characteristics of Mutual Funds

Mutual fund stock holdings data are a natural source for constructing networks, as they yield insight into the stock selection decisions of funds over time. While portfolio holdings can reveal a fund manager's preference, joint stock holdings can capture the connections between funds. When there are multiple funds investing in the same stock with substantial weights, it is likely that some common belief, knowledge, or information is shared among those fund managers, which forms a group of funds. We argue that mutual funds' grouping structure can be used to reveal the extent of fund access to market information.

3.2.1 Network of Mutual Funds

The network of mutual funds is constructed based on joint stock holdings among funds. The nodes in this network are mutual funds. A linkage between two funds is established if they hold the same stock with substantial weights in their own portfolios (over 5% of their respective portfolios in terms of market value). Figure 3.1 shows an example of stock holdings from four mutual funds, where the four circles are funds, and the five rectangles are stocks. This reflects the raw data on mutual fund stock holdings from which a network of funds can be constructed. In Figure 3.2, the corresponding network of these four mutual funds is plotted, where the disconnection between fund a and fund b indicates that there is no commonality in stock holdings between them.

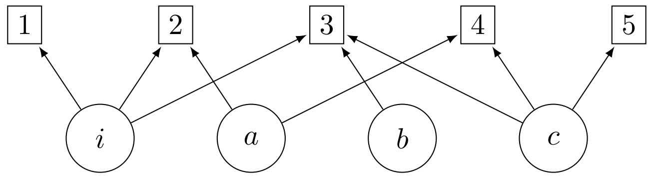


Figure 3.1: **Example of Mutual Fund Stock Holdings**
Four mutual funds i, a, b, and c are investing in stocks 1, 2, 3, 4, and 5. The arrows indicate each fund's stock holdings. For example, fund i is holding stocks 1, 2, and 3.

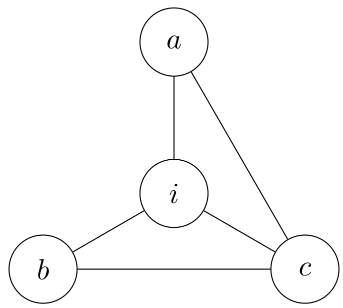


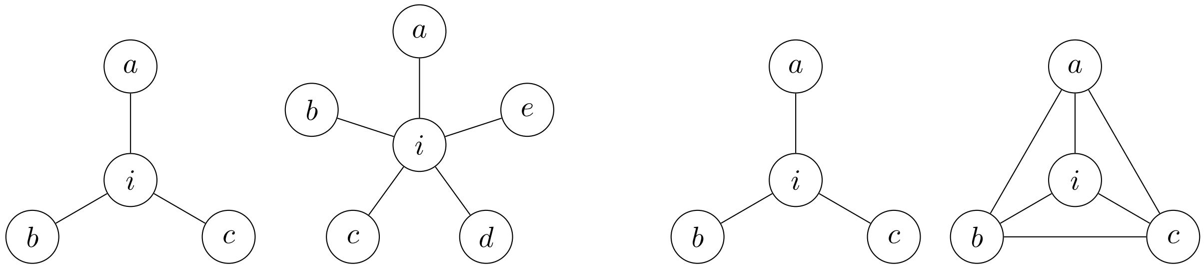
Figure 3.2: **Example of a Network of Mutual Funds**
There are four mutual funds: i, a, b, and c. A link between two funds is established if there exists at least one stock held by both funds. This network corresponds to the stock holdings depicted in Figure 3.1. For example, fund a is not connected to fund b, as there is no commonality in stock holdings between fund a and fund b.

3.2.2 Network Coefficients

Degree and clustering coefficient are commonly used in network studies. In a network, the degree coefficient (D_i) for a node i is defined as the number of *direct links* that node i has.[27] In our context, it describes the number of funds that are holding the same stocks that fund i does. The clustering coefficient (C_i) for a node i is defined as

$$C_i = \frac{\text{The number of links among } \mathcal{N}_i}{\text{The number of maximum possible links among } \mathcal{N}_i},$$

where $\mathcal{N}_i$ ("neighbors" of i) is the set of nodes that is connected to node i.[28] That is, the set of funds that holds common beliefs with fund i. It is worth mentioning that the denominator in the expression for C_i is a nonlinear function of D_i, as $D_i \times (D_i - 1)/2$. In general, the more direct links (or neighbors) a fund has, the more links can potentially be found among them. In graph theory, the clustering coefficient captures the extent to which nodes in a graph tend to cluster together. In our context, it describes the closeness among the funds to which fund i is connected. Figure 3.3 provides sample calculations of the two coefficients for fund i.



$D_i = 3$ $D_i = 5$ $C_i = \frac{0}{3} = 0$ $C_i = \frac{3}{3} = 1$

Figure 3.3: **Examples of Degree and Clustering Coefficients**
The degree coefficient (D_i) of fund i is calculated by counting the number of nodes that are connected to fund i. The clustering coefficient (C_i) of fund i is calculated by using the actual number of links among the neighboring funds a, b, and c divided by the maximum possible links among them, which can be calculated as
$D_i \times (D_i - 1)/2 = 3 \times (3 - 1)/2 = 3$.

Degree and clustering coefficients are informative in studying a mutual fund's investment style. They reveal the stock selection preference of a fund manager. If a fund has a large value of D_i, many other funds are currently making the same stock holding decisions as this fund. This usually happens when funds share market information sources that are very influential, for example, research reports produced by well-known research institutions.

The clustering coefficients, however, reveal the closeness among the funds that are connected to fund i. If we define the group as many funds that invest in the same stock, C_i will indicate the "location" of a fund relative to different groups of other funds. On the one hand, a fund with a higher value of C_i is located in the center of some group, as other funds that are connected to fund i are also connected to each other, which defines this group (see Figure 3.4). On the other hand, a fund with a lower value of C_i is located between different groups (instead of in the center of any one group). It may have connections with funds from different groups (see Figure 3.5).

In explaining mutual fund performance persistence, we pay more attention to the clustering coefficient, as it is closely related to a fund's information characteristics. The intuition is that a lower clustering level indicates a higher level of diversification in the market information sources to which a fund has access. Holding the degree coefficient constant, the value of C_i captures the independence of fund i's information sources. If C_i takes the maximum value of one, fund i has knowledge of only one source (or one aspect of the market). It is unlikely for fund i to obtain a complete information about the market. For example, in Figure 3.3, fund i has $D_i = 3$ in the 3rd and 4th scenarios, but in the 3rd scenario (with $C_i = 0$), fund i is more likely to obtain complete market information because a, b, and c are from three unconnected groups representing three different aspects of the complete market status, whereas in the 4th scenario (with $C_i = 1$) a, b, and c form a single group. We would expect that funds with lower values of C are more likely to have persistent performance.

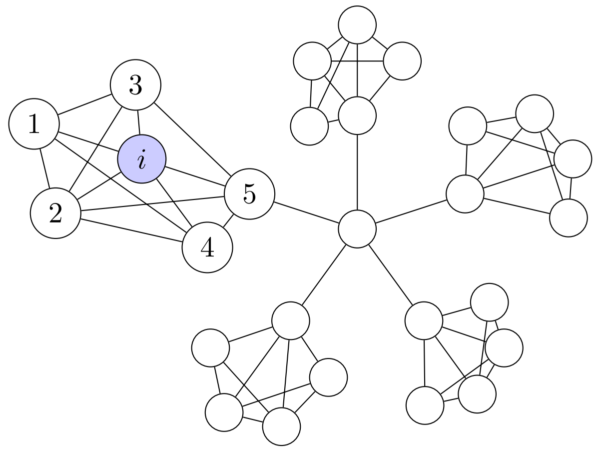


Figure 3.4: **Example of a Network Where Fund I Is Located Within a Single Group of Funds**

With $D_i = 5$ and $C_i = 8/10 = 0.8$ (8 links among fund i's five neighbors can potentially form of $5 \times (5-1)/2 = 10$ links at most), fund i is in the center of a group of funds, and it has no connections to funds in other groups.

3.3 Empirical Study

According to our hypothesis, mutual funds have different strengths in generating persistent performance because fund managers' skills are constrained by the market information that is available to them. The completeness of information to which a fund has access can be captured by its stock holdings and its relationship to other funds. Our empirical study is conducted in four steps: i) we construct a network of funds at each point in time based on their portfolio holdings; ii) we calculate the two indices (degree and clustering) for the funds, representing their information accessibility; iii) we estimate fund performance (past and future) over time based on common benchmark models; and iv) given the panel of mutual funds with their performance and the two network-based indices, we apply a panel regression analysis and tests based on the sorted portfolio method to examine the effects of both indices on the persistence of fund performance.

3.3.1 Data

The data are collected from the CRSP U.S. Mutual Fund Database and Kenneth French's data library. Because we want to utilize stock holdings to construct networks, the sample of mutual funds is limited to equity funds, excluding equity index funds.[29] The network of mutual funds is constructed for every month using data on funds' stock holdings, where a link between two funds is established if they hold the same stock at over 5% of their respective portfolios in terms of market value.[30] At time t, a network of funds is constructed based on stock holdings. For each fund i, two network coefficients, $D_{i,t}$ and $C_{i,t}$, are calculated as described in Section 3.2. Mutual funds' monthly returns are collected from the fund returns

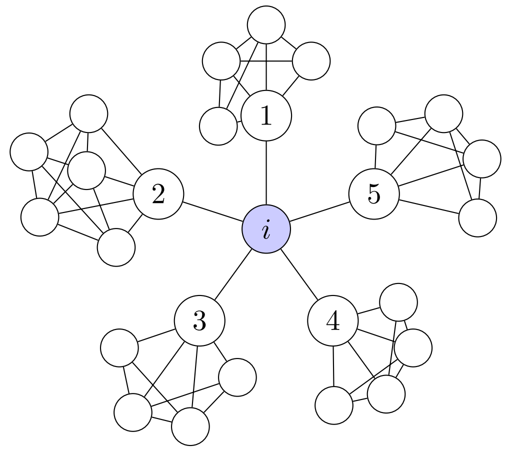


Figure 3.5: **Example of a Network Where Fund I Is Connected to Different Groups of Funds**
With $D_i = 5$ and $C_i = 0$, fund i is not in the center of any group but has connections to all five groups.

dataset. Monthly returns are calculated including reinvested dividends from one period to the next and excluding all management expenses and 12b fees.[31] The time series of risk factor returns are collected from Kenneth French's data library. Our sample period is from January 31, 2001 to September 30, 2014. The observation frequency is monthly.

To examine performance persistence, we need to measure past and future performance for each fund in each month, and we use $\alpha_{i,t}^{pre}$ and $\alpha_{i,t}^{post}$ to denote these two measures for fund i at time t, respectively. The superscripts correspond to the fact that they are estimated by using the pre- and post-period returns from the perspective of time t. For example, with a 24-month investment horizon, the pre-period performance at time t is based on mutual fund's returns from $(t-23)$ to t, while the post-period performance is based on returns from $(t+1)$ to $(t+24)$. In particular, the alphas are the estimated intercepts in time series regressions of fund excess returns on the benchmark factor returns.[32] They are considered the abnormal returns that are not explained by the benchmark model and are thus used as performance measures. We use the CAPM and Fama-French Three-Factor model as the benchmark models

$$r_{i,\tau} = \alpha_{i,T} + \beta_{i,T} RMRF_\tau + e_{i,\tau} \qquad \tau = 1, 2, 3, ..., T \tag{3.1}$$

$$r_{i,\tau} = \alpha_{i,T} + \beta_{i,T}^m RMRF_\tau + \beta_{i,T}^s SMB_\tau + \beta_{i,T}^b HML_\tau + e_{i,\tau} \qquad \tau = 1, 2, 3, ..., T \tag{3.2}$$

where $r_{i,\tau}$ is the mutual fund monthly return in excess of the one-month T-bill return; RMRF is the excess return on a value-weighted market proxy; and SMB and HML are returns on factor-mimicking portfolios for size and book-to-market equity. The intercept, $\alpha_{i,T}$, is estimated by the OLS regression.

It is worth mentioning that the stock holdings dataset is indexed by the portfolio code, while the fund return dataset is indexed by fund code, and the mapping between these two codes is not one-to-one. For example, a fund code may correspond to more than one portfolio code on different dates because the portfolio code for the same fund is updated over time. Moreover, a portfolio code may correspond to multiple fund codes on the same date for a different reason: a fund company usually offers more than one class of fund[33] at the same time, which each have separate fund codes but are invested in the same portfolio of stocks (and directed by the same managers) and thus have the same portfolio code. To address this, we screen the data by including only one representative fund for each portfolio. In particular, we pick the fund that has the highest value of total net asset (TNA) over the sample period. In this way, each portfolio code in our sample will correspond only to the fund with the highest TNA over the fund's life.

Table 3.1 contains the summary statistics for the final sample. The period is January 2001 to September 2014, and the sample frequency is monthly. The panel of mutual funds consists of 5,796 funds and 86,480 monthly observations for C and D. Our panel is unbalanced: the number of observations for each firm varies between 1 and 128.

3.3.2 Autoregressive Model

Given the panel of mutual funds, we test performance persistence by conducting a regression analysis based on the equation

$$\begin{aligned}\alpha_{i,t}^{post} = \beta_0 + \beta_1 \alpha_{i,t}^{pre} + \beta_2 C_{i,t} + \beta_3 D_{i,t} + \beta_4 C_{i,t} D_{i,t} \\ + \left[\beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t}\right] \alpha_{i,t}^{pre} + \epsilon_{i,t},\end{aligned} \tag{3.3}$$

where $\alpha_{i,t}^{pre}$ and $\alpha_{i,t}^{post}$ are mutual fund i's pre- and post-period risk-adjusted performance at time t estimated from one of the two benchmark models (see Equations 3.1 and 3.2); $C_{i,t}$ and $D_{i,t}$ are the clustering and degree coefficients for fund i measured at time t standardized to have zero mean and unit standard deviation for ease of interpretation. Notice that, by definition, the denominator in C's expression is a nonlinear function of D. It is better to disentangle the two indices, as we have both variables in the regression equation. Thus, we make an adjustment to the clustering coefficient (and call it the adjusted clustering coefficient), taking C as the number of links among a fund's neighbors, which is the numerator in the original expression of C. The implication remains the same. The partial effects of the network coefficients on performance persistence are captured by $\{\beta_5, \beta_6, \beta_7\}$. In particular, according to our argument, we would expect a negative value of $\hat{\beta}_5$, which indicates that a lower clustering level helps generate persistent performance, holding other things constant.

The estimated the partial effects of network coefficients are presented in Table 3.2, using the Fama-French Three-Factor model as the performance benchmark and taking 24 months as the investment horizon. We examine the results under eight different combinations of restrictions: the fund individual fixed effect (in all cases), the time effect, the robustness in standard error, and the estimation method (Feasible GLS vs OLS). The fixed effect is controlled by taking the first-order difference of all variables.[34] We can see that in all cases the regression coefficient on the interaction term between α^{pre} and C is significantly less than zero ($\hat{\beta}_5 < 0$). This indicates that the cluster level has a negative effect on the strength of persistence. This is consistent with our argument about information accessibility. With a lower value of C_i, the fund i is connected to the funds that are themselves separated. Such diversified investment can be understood as the versatility of fund i's information sources. Therefore, the performance of a fund with a low C_i will mostly depend on its manager's skill because the constraint on information sources is weak, while a high C_i fund's performance is attributed more to luck, which is less likely to persist over time. The coefficient on the three-way interaction among α^{pre}, C, and D is significantly greater than zero ($\hat{\beta}_7 > 0$), which indicates that such a negative effect is diminishing as the level of the degree coefficient increases. The intuition is that adding or removing a link among the funds that are connected to fund i is of little importance if there are several connected funds.[35]

The size of the estimated coefficients ($\hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7$) determines the difference in persistence strength among different groups of funds captured by indices C and D. Thus, for these two indices to be meaningful in differentiating funds, the difference in persistence should be detectable. We test this in Table 3.4. First, we choose a high and a low value for C and D as their 90th and 10th percentiles, respectively. Then, for each such combination of two values of C and D, we test the null hypothesis that the overall effects of the two network indices is equal to zero. That is,

$$H_0 : \qquad \text{Overall Effect} \equiv \beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t} = 0,$$

which is based on the estimation results from the main regression (as in Table 3.2) and specific values of $C_{i,t}$ and $D_{i,t}$. Using the results from the last model specification in Table 3.2, a statistically significant positive effect exists only for funds with low values of both C and D (with a p-value less than 0.001). And the test on marginal effects also confirms that the clustering coefficient has a negative effect on generating performance persistence (with a p-value equal to 0.035), holding the degree coefficient at its low value. We can see that the results also suggest that a negative overall effect exists for funds with high values of both C and D, but the marginal effects of either are not significant if holding the other at a high value. Having said that, the signs of the coefficients are still in the right direction. For example, with a high value of D, the marginal effect of C is still negative but with a smaller absolute value of the difference in the overall effect and a slightly greater p-value (0.093) compared with the low-D case. This confirms that the effect of C is diminishing in D.

As robustness checks, we conduct the same regression analysis and associated tests using different benchmark models and investment horizons. The results are robust to the use of the Fama-French Three-Factor model and the CAPM with longer investment horizons (three and four years). These regression results are reported in Table 3.5. As an illustration, in Table 3.6, we report the test results for the case of the Fama-French Three-Factor model with three-year horizon.

Another interesting finding is reported in Table 3.7: the pattern of results becomes stronger, in terms of significance level, if we restrict the sample period to 2012. It is the latest year that we can use because the fund return data are available only until September, 2014 and we need at least 18 months[36] to estimate future performance.

3.3.3 Sorted Portfolio Methods

The persistence test based on sorted portfolios is another commonly used method of studying mutual fund performance persistence, which does not rely on a regression model. The general procedure is as follows: first, on each date, the mutual funds are sorted into several portfolios based on their past performance ($\alpha_{i,t}^{pre}$); those portfolios' future performance can be calculated, as we know the composition of each portfolio and the fund performance is available ($\alpha_{i,t}^{post}$); over time, we calculate a sequence of performance for each sorted portfolio; simple t-tests are conducted for portfolios, and a significant performance spread between high-rank and low-rank portfolios is considered evidence of persistent performance. We conduct this analysis for our sample and obtain similar results as obtained through the autoregressive model.

For comparison with the results in Table 3.2, we use the same benchmark model (the Fama-French Three-Factor Model) and investment horizon (24 months) to estimate the performance. At time t, using the past performance ($\alpha_{i,t}^{pre}$), we sort the mutual funds into above-median (Top 50%) and below-median (Bottom 50%) portfolios whose future performance can be calculated as the average fund future performance ($\alpha_{i,t}^{post}$) over the funds in each portfolio. The spread is the difference between the two portfolios. As reported in the Panel A of Table 3.8, the below-median portfolio has a negative post-ranking abnormal returns on average (-0.0007), with a p-value equal to 0.003. Although the average post-ranking performance from the above-median portfolio is not statistically significantly different from zero, the spread between the top and the bottom is significantly greater than zero (with a p-value equal to 0.0019). This suggests that, on average, fund performance is persistent, and one can make arbitrage profits (0.07%) from a momentum strategy — taking long positions in past winner funds and short positions in loser funds.

To study heterogeneity in the strength of persistent, we further sort funds according to their network indices. The variables *High Value* and *Low Value* correspond to funds with above-median and below-median value of $C_{i,t}$ and $D_{i,t}$ respectively. *All* includes all funds under the corresponding criterion. From these three classes of C and D, we define nine groups of funds. For each group, we perform the test by the sorted portfolio method described above and report the *Top-Bottom Spread* (and p-value) for each group.

We find that the low-C and low-D group contributes the most to the overall persistence among all funds, which has the largest *Top-Bottom spread* with a p-value less than 0.001. In other words, there will be a considerable increase in profit (0.18% vs 0.07%) from the momentum strategy if one can exploit the information contained in the fund network. The test of the marginal effect confirms that the clustering coefficient has a negative effect on generating momentum, as there is a 0.12% loss in profit from the risk-free momentum strategy if one uses only the funds with high-C.

3.4 Simulation Study

The purpose of this simulation analysis is to assess the relevance and appropriateness of the two network indices in representing the unobserved factor, information accessibility (IA), which is conveniently known in our simulated environment.

3.4.1 Experiment Design

The simulation considers a simple setting in which N fund managers make investment decisions, searching for the best from M stocks to hold, according to fund characteristics, of their skill and IA. Based on the stock holdings, the network of mutual funds is constructed, and network coefficients C and D for N funds are recorded. Then, the relationships between IA and the network coefficients are examined.

In particular, each of the N funds is fully characterized by its skill and information accessibility level. The M stocks are ranked from 1 to M. Each fund will choose a certain number of stocks to hold from its own *stock observation set*, which is a subset of the M stocks. IA_i is the information parameter that determines the size of the stock observation set for fund i, which is the number of stocks about which fund i has knowledge. $Skill_i$ is the skill parameter ranging between 0 and 1, indicating the probability that fund i will correctly rank the stocks in its observation set and make the investment decisions by picking the high-rank stocks. For example, for a fund with $Skill_i = 0.5$, there is 50% chance that fund i will correctly find the best stocks to hold, while with $Skill_i = 0$, fund i will randomly rank the stocks in its observation set and make the holding decision.

The simulated sample is based on an experiment design with 2,400 stocks and 1,600 funds with 40 levels of $Skill_i$ and 40 levels of IA_i.[37] The experiment has 10 rounds. Each fund chooses 3 stocks in each round according to its rank. Based on the stock holdings in each round, the network of mutual funds is constructed, and the network coefficients C_i and D_i for fund i are recorded.

Figure 3.6 provides the comparison between the simulated sample and the real sample that we use in the empirical study presented Section 3.3 in terms of the distribution of C_i and D_i.

3.4.2 Results

In Table 3.9, we regress C on fund characteristics to test the hypothesis that a fund with more complete information (high IA) connects to funds that are diversified (low C), holding other things constant. The estimated coefficient on IA is significantly less than zero with a p-value less than 0.001. In the specifications that include squared terms and/or interaction terms of $\{IA, D, Skill\}$, the effect of $Skill$ on C disappears, as expected, because we believe that C represents an information characteristic rather than skill.

In Table 3.10, we examine the determinants of information accessibility by regressing IA on the two network indices. We obtain a consistent result that C has a negative effect on IA among the specifications that include both C and D, which suggests that we should include both indices in empirical studies (as in Equation 3.3), as this allows us to observe the partial effect of C on IA when holding D constant.

The simulation results suggest that our indices, C and D, can capture information accessibility and that the clustering coefficient and information access are negatively associated with each other, holding the degree coefficient constant. We take this as an evidence supporting the use of C and D as proxies for information accessibility in the empirical study and for interpreting a low value C as a sign that the fund has more complete information.

3.5 Conclusions

In this paper, we propose a new explanation for mutual fund performance persistence. We argue that fund managers' stock selection skills are constrained by the market information that is available to them. If all funds can make investment decisions based on the same market information, a high-skill manager will have performance that constantly dominates others, provided that skills do not change over a short period. If a fund has only incomplete information about the market, its performance will not be consistent with manager skill, and it could outperform or underperform with respect to the true skill level. Therefore, its performance will not persist. To identify the extent of a mutual fund access to market information, we construct a network of funds based on their common holdings of stocks and use two network coefficients as the proxies. Based on the sample of equity funds included in the CRSP U.S. Mutual Fund Database from January 31, 2001 to September 31, 2014, we find that a mutual fund with more complete information about the market is more likely to show persistence in its performance.

3.6 Notes

[24]The terms *momentum* and *persistence* are interchangeable in literature.

[25]Other examples can be found in Simon (1971), Kahneman (1973), Barber and Odean (2008), Veldkamp (2006), Kacperczyk et al. (2009), and Gupta-Mukherjee and Pareek (2014). Studies that argue or show that information is related to fund performance include Dybvig and Ross (1985), Coval and Moskowitz (1999), Coval and Moskowitz (2001), Nanda, Wang, and Zheng (2004), Kacperczyk, Sialm, and Zheng (2005), and Cohen, Frazzini, and Malloy (2008).

[26]Equity funds represented 52% of U.S. mutual fund assets as of 2013 (Investment Company Institute, 2014).

[27]$D_i = 0, 1, 2, ..., N-1$, where N is the total number of funds in a network.

[28]$C_i \in [0, 1]$. The maximum value of 1 occurs when all funds that are connected to fund i are connected among themselves, whereas the minimum value of 0 occurs when the funds that are connected to fund i are themselves separated.

[29]This is accomplished by including the funds with *Lipper Asset Code* = "Equity Funds", excluding the funds with *Index Fund Identifier* = {" Index-based fund", "Pure Index fund", or "Index fund enhanced"}, and excluding the funds with *Lipper Objective Code* = "S&P 500 Index Objective Funds".

[30]Using a threshold will make the sample more representative and rule out the undesired cases in which funds hold (or try to replicate) a market index in their portfolio.

[31]*12b fees* are annual marketing or distribution fees on mutual funds.

[32]To address estimation noise due to the missing observations in fund return, we required that the alpha for fund i is estimated when at least half of fund i's return observations are not missing.

[33]For example, the institutional class fund and retail class fund from the same company are with same portfolio code but different fund codes.

[34]The conventional within estimator is biased due to endogeneity.

[35]One interesting extension is presented in Table 3.3, in which the risk-adjusted performance measure, $\alpha_{i,t}$ as in Table 3.2, is replaced by systematic risk measure, $Beta_{i,t}$, the CAPM Beta. Result suggests that network coefficients can also be applied in systematic risk analysis, which is one of the topics for future study.

[36]This is the half length of a 36-month horizon.

[37]$Skill_i$ is evenly distributed between 0 and 1. IA_i is evenly distributed between 3 and 2,400.

3.7 Tables and Figure

Table 3.1: **Summary Statistics**

This table presents summary statistics for the main variables in our sample. The data cover the period from January 2001 to September 2014 with monthly frequency. The network coefficients C and D are calculated based on the structure of the fund network in each month. The performance measures are the estimated intercepts in time series regressions of fund excess returns on the benchmark factor returns. We use the CAPM and the Fama-French Three-Factor model (*FF*) as the benchmark models (see Equations 3.1 and 3.2) with one- to five-year investment horizons to estimate α. For each month, α^{pre} and α^{post} are estimated based on the returns in pre- and post-period from that month's perspective. The panel of mutual funds consists of 5,796 funds and 86,480 monthly observations for C and D. The observation number for the alphas varies because it takes more monthly return data to estimate an alpha with a longer investment horizon. Our panel is unbalanced: the number of observations for each firm varies between 1 and 128.

			Mean	**Std. Dev.**	**Min.**	**Max.**	**N**
C			0.850441	0.218744	0	1	86480
D			51.79667	103.353332	2	632	86480
CAPM	1-year	α^{pre}	-0.000841	0.007601	-0.134278	0.068729	83297
		α^{post}	-0.001257	0.007406	-0.083497	0.065989	73170
	2-year	α^{pre}	-0.000653	0.005826	-0.092903	0.054407	78860
		α^{post}	-0.001435	0.005669	-0.066232	0.046738	59321
	3-year	α^{pre}	-0.000402	0.005117	-0.076604	0.048361	74202
		α^{post}	-0.001609	0.005199	-0.051773	0.030576	47251
	4-year	α^{pre}	0.000049	0.004541	-0.059859	0.048361	68804
		α^{post}	-0.001579	0.004977	-0.054186	0.030971	34746
	5-year	α^{pre}	0.000213	0.004179	-0.053255	0.047561	62485
		α^{post}	-0.001683	0.005134	-0.054186	0.027031	22793
FF	1-year	α^{pre}	-0.00193	0.00888	-0.146465	0.265407	83297
		α^{post}	-0.002018	0.008738	-0.119498	0.071904	73170
	2-year	α^{pre}	-0.00096	0.006004	-0.111639	0.066763	78860
		α^{post}	-0.001342	0.005747	-0.066887	0.050439	59321
	3-year	α^{pre}	-0.000644	0.005047	-0.087451	0.05219	74202
		α^{post}	-0.001538	0.005179	-0.045517	0.036395	47251
	4-year	α^{pre}	-0.000259	0.004382	-0.062898	0.0366	68804
		α^{post}	-0.001397	0.004839	-0.046879	0.028854	34746
	5-year	α^{pre}	-0.00006	0.003905	-0.057316	0.032507	62485
		α^{post}	-0.001532	0.004874	-0.046879	0.02943	22793

Table 3.2: **Partial Effects of Network Coefficients on Performance Persistence**

Partial effects are estimated with various considerations for the model

$$\alpha_{i,t}^{post} = \beta_0 + \beta_1 \alpha_{i,t}^{pre} + \beta_2 C_{i,t} + \beta_3 D_{i,t} + \beta_4 C_{i,t} D_{i,t} + \left[\beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t}\right] \alpha_{i,t}^{pre} + \epsilon_{i,t},$$

where $\alpha_{i,t}^{pre}$ and $\alpha_{i,t}^{post}$ are mutual fund i's ranking period and post-ranking period risk-adjusted performance at time t, respectively, estimated from the Fama-French Three-Factor model with an investment horizon of 24 months. The ranking period performance at time t is based on mutual fund's returns from $(t-23)$ to t, while the post-ranking period performance is based on returns from $(t+1)$ to $(t+24)$. $C_{i,t}$ and $D_{i,t}$ are the adjusted clustering and degree coefficients for fund i at time t standardized to have a zero mean and unit standard deviation. The partial effects of network coefficients on performance persistence are captured by $\{\beta_5, \beta_6, \beta_7\}$. The last four rows indicate whether a particular issue in estimation is considered for each regression. The sample consists of 2,061 mutual funds. The sample period is from January 1, 2001 to September 30, 2014. The observation frequency is monthly.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
C	β_2	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001
D	β_3	−0.0002∗	−0.0002∗	−0.0002∗	−0.0002∗	−0.0002∗	−0.0002∗	−0.0002∗	−0.0002∗
$C \times D$	β_4	−0.0000	−0.0000	−0.0000	−0.0000	0.0000	0.0000	0.0000	0.0000
$\alpha^{pre} \times C$	β_5	−0.1001∗∗∗	−0.0995∗∗∗	−0.1001∗∗	−0.0995∗	−0.0795∗∗	−0.0795∗∗	−0.0795∗	−0.0795∗
$\alpha^{pre} \times D$	β_6	0.0257	0.0256	0.0257	0.0256	0.0172	0.0172	0.0172	0.0172
$\alpha^{pre} \times C \times D$	β_7	0.0160∗∗∗	0.0159∗∗∗	0.0160∗∗∗	0.0159∗∗∗	0.0133∗∗∗	0.0133∗∗∗	0.0133∗∗	0.0133∗∗
Fixed Effect		*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Time Effect		*No*	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Robust SE		*No*	*No*	*Yes*	*Yes*	*No*	*No*	*Yes*	*Yes*
FGLS/OLS		*OLS*	*FGLS*	*OLS*	*FGLS*	*OLS*	*FGLS*	*OLS*	*FGLS*

∗ $p < 0.05$; ∗∗ $p < 0.01$; ∗∗∗ $p < 0.001$

Table 3.3: **Partial Effects of Network Coefficients on CAPM Beta Persistence**

Partial effects are estimated with various considerations for the model

$$Beta_{i,t}^{post} = \beta_0 + \beta_1 Beta_{i,t}^{pre} + \beta_2 C_{i,t} + \beta_3 D_{i,t} + \beta_4 C_{i,t} D_{i,t} + \left[\beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t}\right] Beta_{i,t}^{pre} + \epsilon_{i,t},$$

where $Beta_{i,t}^{pre}$ and $Beta_{i,t}^{post}$ are mutual fund i's ranking period and post-ranking period CAPM Beta at time t, respectively, estimated from the CAPM model with an investment horizon of 24 months. The ranking period CAPM Beta at time t is based on mutual fund's returns from $(t-23)$ to t, while the post-ranking period CAPM Beta is based on returns from $(t+1)$ to $(t+24)$. $C_{i,t}$ and $D_{i,t}$ are the adjusted clustering and degree coefficients for fund i at time t standardized to have a zero mean and unit standard deviation. The partial effects of network coefficients on CAPM Beta persistence are captured by $\{\beta_5, \beta_6, \beta_7\}$. The last four rows indicate whether a particular issue in estimation is considered for each regression. The sample consists of 2,061 mutual funds. The sample period is from January 1, 2001 to September 30, 2014. The observation frequency is monthly.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
C	β_2	0.1276∗∗∗	0.1269∗∗∗	0.1276∗∗∗	0.1269∗∗∗	0.0955∗∗∗	0.0955∗∗∗	0.0955∗∗∗	0.0955∗∗∗
D	β_3	−0.0563∗∗∗	−0.0566∗∗∗	−0.0563∗∗∗	−0.0566∗∗∗	−0.0562∗∗∗	−0.0562∗∗∗	−0.0562∗∗∗	−0.0562∗∗∗
$C \times D$	β_4	−0.0169∗∗∗	−0.0168∗∗∗	−0.0169∗∗∗	−0.0168∗∗∗	−0.0120∗∗∗	−0.0120∗∗∗	−0.0120∗∗∗	−0.0120∗∗∗
$Beta^{pre} \times C$	β_5	−0.1455∗∗∗	−0.1448∗∗∗	−0.1455∗∗∗	−0.1448∗∗∗	−0.1080∗∗∗	−0.1080∗∗∗	−0.1080∗∗∗	−0.1080∗∗∗
$Beta^{pre} \times D$	β_6	0.0705∗∗∗	0.0708∗∗∗	0.0705∗∗∗	0.0708∗∗∗	0.0658∗∗∗	0.0658∗∗∗	0.0658∗∗∗	0.0658∗∗∗
$Beta^{pre} \times C \times D$	β_7	0.0189∗∗∗	0.0188∗∗∗	0.0189∗∗∗	0.0188∗∗∗	0.0133∗∗∗	0.0133∗∗∗	0.0133∗∗∗	0.0133∗∗∗
Fixed Effect		*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Time Effect		*No*	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Robust SE		*No*	*No*	*Yes*	*Yes*	*No*	*No*	*Yes*	*Yes*
FGLS/OLS		*OLS*	*FGLS*	*OLS*	*FGLS*	*OLS*	*FGLS*	*OLS*	*FGLS*

∗ $p < 0.05$; ∗∗ $p < 0.01$; ∗∗∗ $p < 0.001$

Table 3.4: **Tests of Persistence Strength for Different Values of C and D: Fama-French Three-Factor Model, 24-Month Horizon**

The table lists the joint tests of the overall effects of $C_{i,t}$ and $D_{i,t}$ on performance persistence with different values of $C_{i,t}$ and $D_{i,t}$.

$$H_0: \qquad \text{Overall Effect} \equiv \beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t} = 0,$$

where $\{\hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7\}$ are estimated according to the last model specification in Table 3.2.

$$\alpha_{i,t}^{post} = \beta_0 + \beta_1 \alpha_{i,t}^{pre} + \beta_2 C_{i,t} + \beta_3 D_{i,t} + \beta_4 C_{i,t} D_{i,t} + \left[\beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t}\right] \alpha_{i,t}^{pre} + \epsilon_{i,t},$$

where $\hat{\beta}$s are estimated by FGLS controlling for fixed effects and time effects with robust standard errors. The variables and notation are detailed in the caption of Table 3.2. *High Value* and *Low Value* correspond to the 90th and 10th percentile of $C_{i,t}$ and $D_{i,t}$ respectively. The values of $High{-}Low$ are the marginal effects of $C_{i,t}$ or $D_{i,t}$ when moving from a low level to high level, holding the other variable constant at either a high or low level. Performance is estimated from the Fama-French Three-Factor model with an investment horizon of 24 months. The sample and variable notation are the same as in Table 3.2. The numbers in parentheses are p-values.

		C		
		High Value	*Low Value*	$High - Low$
D	*High Value*	−0.0289∗	0.0559	−0.0849
		(0.049)	(0.365)	(0.093)
	Low Value	−0.1377	0.0171∗∗∗	−0.1548∗
		(0.052)	(0.000)	(0.035)
	$High - Low$	0.1087	0.0388	
		(0.182)	(0.515)	

∗ $p < 0.05$; ∗∗ $p < 0.01$; ∗∗∗ $p < 0.001$

Table 3.5: **Robustness by Performance Benchmark and Investment Horizon**

Partial effects are estimated with different performance benchmarks and investment horizons for the model

$$\alpha_{i,t}^{post} = \beta_0 + \beta_1 \alpha_{i,t}^{pre} + \beta_2 C_{i,t} + \beta_3 D_{i,t} + \beta_4 C_{i,t} D_{i,t} + \left[\beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t}\right] \alpha_{i,t}^{pre} + \epsilon_{i,t},$$

where $\alpha_{i,t}^{pre}$ and $\alpha_{i,t}^{post}$ are mutual fund i's ranking period and post-ranking period risk-adjusted performance at time t, respectively, estimated from the Fama-French Three-Factor model or the CAPM model with investment horizon of 36 or 48 months. The ranking period performance with a 36-month horizon at time t is based on mutual fund's returns from period $(t-35)$ to t, while the post-ranking period performance is based on returns from $(t+1)$ to $(t+36)$. $C_{i,t}$ and $D_{i,t}$ are the adjusted clustering and degree coefficients for fund i at time t, standardized to have zero mean and unit standard deviation. The partial effects of the network coefficients on performance persistence are captured by $\{\beta_5, \beta_6, \beta_7\}$. In all cases, the $\hat{\beta}$s are estimated by FGLS controlling for fixed effects and time effects with robust standard errors. The sample consists of 2,061 mutual funds. The sample period is from January 1, 2001 to September 30, 2014. The observation frequency is monthly.

		Fama-French Three-Factor		*CAPM*	
		36 months	48 months	36 months	48 months
C	β_2	0.0003∗∗∗	0.0000	0.0006∗∗∗	−0.0000
D	β_3	−0.0002∗∗∗	0.0000	−0.0003∗∗∗	0.0000
$C \times D$	β_4	−0.0000∗∗∗	−0.0000	−0.0001∗∗∗	−0.0000
$\alpha^{pre} \times C$	β_5	−0.0724∗∗	−0.0733∗	−0.0692∗	−0.0736∗∗
$\alpha^{pre} \times D$	β_6	0.0140	0.0135	0.0246	0.0342∗
$\alpha^{pre} \times C \times D$	β_7	0.0116∗∗∗	0.0111∗	0.0106∗∗	0.0086∗
Fixed Effect		*Yes*	*Yes*	*Yes*	*Yes*
Time Effect		*Yes*	*Yes*	*Yes*	*Yes*
Robust SE		*Yes*	*Yes*	*Yes*	*Yes*
FGLS/OLS		*FGLS*	*FGLS*	*FGLS*	*FGLS*

∗ $p < 0.05$; ∗∗ $p < 0.01$; ∗∗∗ $p < 0.001$

Table 3.6: **Tests of Persistence Strength with Different Values of C and D: Fama-French Three-Factor Model, 36-Month Horizon**

The table lists the joint tests of the overall effects of $C_{i,t}$ and $D_{i,t}$ on performance persistence with different values of $C_{i,t}$ and $D_{i,t}$.

$$H_0: \qquad \text{Overall Effect} \equiv \beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t} = 0$$

where $\{\hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7\}$ are estimated according to the last model specification in Table 3.2

$$\begin{aligned}\alpha_{i,t}^{post} &= \beta_0 + \beta_1 \alpha_{i,t}^{pre} + \beta_2 C_{i,t} + \beta_3 D_{i,t} + \beta_4 C_{i,t} D_{i,t} \\ &\quad + [\beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t}]\, \alpha_{i,t}^{pre} + \epsilon_{i,t},\end{aligned}$$

where $\hat{\beta}$s are estimated by FGLS controlling for fixed effects and time effects with robust standard error. *High Value* and *Low Value* correspond to 90th and 10th percentile of $C_{i,t}$ and $D_{i,t}$, respectively. The values of $High - Low$ are the marginal effects of $C_{i,t}$ or $D_{i,t}$ when moving from low level to high level, holding the other variable constant at either a high or low level. Performance is estimated from the Fama-French Three-Factor model with an investment horizon of 36 months. The sample and variable notation are the same as in Table 3.2. The numbers in parentheses are p-values.

		C		
		High Value	*Low Value*	$High - Low$
D	*High*	−0.0382** (0.005)	0.0502 (0.282)	−0.0884* (0.029)
	Low	−0.1522* (0.012)	0.0162*** (0.000)	−0.1685** (0.008)
	$High - Low$	0.1141 (0.083)	0.0340 (0.454)	

$*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$

Table 3.7: **A Stronger Pattern in 2012**

Partial effects are estimated with different performance benchmarks and investment horizons for the model

$$\alpha_{i,t}^{post} = \beta_0 + \beta_1 \alpha_{i,t}^{pre} + \beta_2 C_{i,t} + \beta_3 D_{i,t} + \beta_4 C_{i,t} D_{i,t} + [\beta_5 C_{i,t} + \beta_6 D_{i,t} + \beta_7 C_{i,t} D_{i,t}]\, \alpha_{i,t}^{pre} + \epsilon_{i,t},$$

where $\alpha_{i,t}^{pre}$ and $\alpha_{i,t}^{post}$ are mutual fund i's ranking period and post-ranking period risk-adjusted performance at time t, respectively, estimated from the Fama-French Three-Factor model or the CAPM model with investment horizons of 12, 24, or 36 months. The ranking period performance with 12 months horizon at time t is based on mutual fund's returns from $(t-11)$ to t, while the post-ranking period performance is based on returns from $(t+1)$ to $(t+12)$. A similar construction applies to the 24- and 36- month investment horizons. $C_{i,t}$ and $D_{i,t}$ are the adjusted clustering and degree coefficients for fund i at time t, standardized to have zero mean and unit standard deviation. The partial effects of network coefficients on performance persistence are captured by $\{\beta_5, \beta_6, \beta_7\}$. In all cases, the $\hat{\beta}$s are estimated by FGLS controlling for fixed effects and time effects with robust standard errors. The sample consists of 1,588 mutual funds. The sample period used in the regression is from January 1, 2012 to December 31, 2012. The data for fund returns before and after 2012 are needed to estimate performance. The observation frequency is monthly.

	Fama-French Three-Factor			*CAPM*		
	12 months	24 months	36 months	12 months	24 months	36 months
C	0.0034***	−0.0004	0.0002	0.0028***	−0.0004*	0.0008***
D	−0.0016***	−0.0004*	−0.0002*	−0.0012***	−0.0005**	−0.0004***
$C \times D$	−0.0004***	0.0001***	−0.0000	−0.0004***	0.0001***	−0.0001***
$\alpha^{pre} \times C$	−0.2256*	−0.2795***	−0.1816***	−0.3025***	−0.2579***	−0.1559***
$\alpha^{pre} \times D$	0.1691**	0.0956***	0.0720***	0.1366***	0.0785*	0.0642**
$\alpha^{pre} \times C \times D$	0.0250*	0.0388***	0.0239***	0.0417***	0.0357***	0.0212***
Fixed Effect	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Time Effect	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Robust SE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
FGLS/OLS	*FGLS*	*FGLS*	*FGLS*	*FGLS*	*FGLS*	*FGLS*

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 3.8: **Portfolios of Mutual Funds Formed on Past Risk-Adjusted Performance: Fama-French Three-Factor Model, 24-Month Horizon**

Mutual funds are sorted in each month from December 30, 2002 to August 31, 2012 into equal-weight top and bottom portfolios based on their past risk-adjusted performance, $\alpha_{i,t}^{pre}$, estimated from the Fama-French Three-Factor model with investment horizon of 24 months—from $(t-23)$ to t. Funds with an above-median $\alpha_{i,t}^{pre}$ comprise *Top 50% Portfolio*, and funds with a below-median $\alpha_{i,t}^{pre}$ comprise *Bottom 50% Portfolio*. Then, each portfolio's performance is estimated during the post-ranking period, from $(t+1)$ to $(t+24)$, from the same model with the same investment horizon.

Panel A lists the top and bottom portfolios' average performance over time. The *Top-Bottom Spread* is the average difference in performance between the two portfolios.

Panel B lists the *Top-Bottom Spread* within different groups of funds. The group is characterized by the values of two network coefficients of funds, $C_{i,t}$ and $D_{i,t}$. *High Value* and *Low Value* correspond to funds with an above-median and a below-median value of $C_{i,t}$ and $D_{i,t}$ respectively. The values of $High - Low$ are the marginal effects of $C_{i,t}$ or $D_{i,t}$ when moving from low level to high level, holding the other variable constant at a certain level. The numbers in parentheses are p-values.

***Panel A**: Risk-Adjusted Performance of Portfolios of Funds*

Top 50% Portfolio:	−0.0000	(0.8735)
Bottom 50% Portfolio:	−0.0007∗∗	(0.0030)
Top-Bottom Spread:	0.0007∗∗	(0.0019)

***Panel B**: Top-Bottom Spreads within Different Groups of Funds*

		C			
		High Value	*Low Value*	*All*	$High - Low$
D	*High Value*	0.0006 (0.1605)	0.0026∗∗ (0.0016)	0.0006 (0.1378)	−0.0020∗ (0.0284)
	Low Value	0.0028∗ (0.0207)	0.0018∗∗∗ (0.0000)	0.0017∗∗∗ (0.0000)	0.0011 (0.3590)
	All	0.0006 (0.1452)	0.0018∗∗∗ (0.0000)	0.0007∗∗ (0.0019)	−0.0012∗∗∗ (0.0000)
	$High - Low$	−0.0023 (0.0832)	0.0008 (0.3713)	−0.0011∗∗∗ (0.0000)	

∗ $p < 0.05$; ∗∗ $p < 0.01$; ∗∗∗ $p < 0.001$

Figure 3.6: **The Distribution of $C_{i,t}$ and $D_{i,t}$ in the Simulated and Real Samples**
The simulated sample is based on the experiment design with 2,400 stocks and 1,600 funds with 40 levels of skills and 40 levels of information accessibility. The experiment has 10 rounds. Each fund chooses three stocks in each round according to its skill and information access characteristics. Based on the stock holdings in each round, the network of mutual funds is constructed, and the network coefficients, $C_{i,t}$ and $D_{i,t}$, for fund i are recorded. The real sample consists of 5,796 mutual funds with the sample period from January 1, 2001 to September 30, 2014. The observation frequency is monthly.

Table 3.9: **Simulation Evidence: Determinants of the Clustering Coefficient**

Regression estimates for various restrictions of the model

$$C_i = \beta_0 + \beta_1 IA_i + \beta_2 D_i + \beta_3 Skill_i + Squared\ terms + Interaction\ terms + \epsilon_i,$$

where C_i and D_i are the clustering and degree coefficients for fund i. IA_i and $Skill_i$ are the information accessibility and the skill parameters, respectively, for fund i as defined in the simulation experiment. IA_i is scaled to a value between 0 and 1 as a proportion measurement relative to the fund that has the highest value of IA_i in the sample. The sample is simulated based on the experiment design with 2,400 stocks and 1,600 funds with 40 levels of skills and 40 levels of information accessibility. The experiment has 10 rounds. Each fund chooses three stocks in each round according to its skill and information access characteristics. Based on the stock holdings in each round, the network of mutual funds is constructed, and the network coefficients C_i and D_i for fund i are recorded. The *Squared terms* and *Interaction terms* indicate whether the regression equation includes the corresponding terms of $\{IA_i, D_i, Skill_i\}$. The numbers in parentheses are p-values obtained by robust standard error estimation.

	(1)	(2)	(3)
IA	−0.0636∗∗∗	−0.2143∗∗∗	−0.1454∗∗∗
	(0.000)	(0.000)	(0.000)
D	0.0009∗∗∗	0.0020∗∗∗	0.0020∗∗∗
	(0.000)	(0.000)	(0.000)
Skill	0.0455∗∗∗	0.0304	0.0455
	(0.000)	(0.295)	(0.153)
Squared terms	*No*	*Yes*	*Yes*
Interaction terms	*No*	*No*	*Yes*
R^2	0.54	0.56	0.56
N	14591	14591	14591

∗ $p < 0.05$; ∗∗ $p < 0.01$; ∗∗∗ $p < 0.001$

Table 3.10: **Simulation Evidence: Determinants of Information Accessibility**

Regression estimates for various restrictions of the model

$$IA_i = \beta_0 + \beta_1 C_i + \beta_2 D_i + \beta_3 C_i D_i + \textit{Squared terms} + \epsilon_i,$$

where C_i and D_i are the clustering and degree coefficients for fund i, respectively. IA_i is the information accessibility parameter for fund i as defined in the simulation experiment, which is scaled to a value between 0 and 1 as a proportion measurement relative to the fund that has the highest value of IA_i in the sample. The sample is simulated based on the experiment design with 2,400 stocks and 1,600 funds with 40 levels of skills and 40 levels of information accessibility. The experiment has 10 rounds. Each fund chooses three stocks in each round according to its skill and information access characteristics. Based on the stock holdings in each round, the network of mutual funds is constructed, and the network coefficients C_i and D_i for fund i are recorded. *Squared terms* indicates whether the regression equation includes the second order terms of C_i and D_i. The numbers in parentheses are p-values obtained by robust standard error estimation.

	(1)	(2)	(3)	(4)	(5)
C	0.1371***	−0.1137***	−0.3176***	−0.1284***	−0.2003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
D		0.0004***	−0.0023***	−0.0047***	−0.0040***
		(0.000)	(0.000)	(0.000)	(0.000)
$C \times D$				0.0058***	0.0028***
				(0.000)	(0.000)
Squared terms	*No*	*No*	*Yes*	*No*	*Yes*
R^2	0.02	0.09	0.23	0.20	0.25
N	14591	14591	14591	14591	14591

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Bibliography

Acemoglu, Daron, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi, 2012, The Network Origins of Aggregate Fluctuations, *Econometrica* 80, 1977–2016.

Allen, Franklin, and Ana Babus, 2008, Networks in Finance, SSRN Scholarly Paper ID 1094883, Social Science Research Network, Rochester, NY.

Andersen, Torben G., Tim Bollerslev, Peter F. Christoffersen, and Francis X. Diebold, 2006, Chapter 15 Volatility and Correlation Forecasting, in *Handbook of Economic Forecasting*, volume 1, 777–878 (Elsevier).

Andersen, Torben G., Tim Bollerslev, and Francis X. Diebold, 2007, Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility, *Review of Economics and Statistics* 89, 701–720.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys, 2003, Modeling and Forecasting Realized Volatility, *Econometrica* 71, 579–625.

AntóN, Miguel, and Christopher Polk, 2014, Connected Stocks, *The Journal of Finance* 69, 1099–1127.

Baker, Malcolm, Lubomir Litov, Jessica A. Wachter, and Jeffrey Wurgler, 2010, Can Mutual Fund Managers Pick Stocks? Evidence From Their Trades Prior to Earnings Announcements, *The Journal of Financial and Quantitative Analysis* 45, pp. 1111–1131.

Barber, Brad M., and Terrance Odean, 2008, All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors, *Review of Financial Studies* 21, 785–818.

Barras, Laurent, Olivier Scaillet, and Russ Wermers, 2010, False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas, *The Journal of Finance* 65, 179–216.

Barroso, Pedro, and Pedro Santa-Clara, 2015, Momentum Has Its Moments, *Journal of Financial Economics* 116, 111–120.

Bollen, Nicolas PB, and Jeffrey A Busse, 2004, Short-Term Persistence in Mutual Fund Performance, *Review of Financial Studies* 18, 569–597.

Brown, Stephen J., and William N. Goetzmann, 1995, Performance Persistence, *The Journal of Finance* 50, 679.

Cao, Jie, and Bing Han, 2013, Cross Section of Option Returns and Idiosyncratic Stock Volatility, *Journal of Financial Economics* 108, 231–249.

Carhart, Mark M., 1997, On Persistence in Mutual Fund Performance, *The Journal of Finance* 52, 57–82.

Chen, Yu-chin, Stephen J. Turnovsky, and Eric Zivot, 2014, Forecasting Inflation Using Commodity Price Aggregates, *Journal of Econometrics* 183, 117–134.

Christiansen, Charlotte, Maik Schmeling, and Andreas Schrimpf, 2012, A Comprehensive Look at Financial Volatility Prediction by Economic Variables, *Journal of Applied Econometrics* 27, 956–977.

Claeskens, Gerda, Jan R. Magnus, Andrey L. Vasnev, and Wendun Wang, 2016, The Forecast Combination Puzzle: A Simple Theoretical Explanation, *International Journal of Forecasting* 32, 754–762.

Clark, Todd E., and Kenneth D. West, 2007, Approximately Normal Tests for Equal Predictive Accuracy in Nested Models, *Journal of Econometrics* 138, 291–311.

Cohen, L, A Frazzini, and C Malloy, 2008, The Small World of Investing: The Use of Social Networks in Bank Decision-Making, *Journal of Political Economy* 116, 951–979.

Cohen, Lauren, and Andrea Frazzini, 2008, Economic Links and Predictable Returns, *The Journal of Finance* 63, 1977–2011.

Cohen, Randolph B, Joshua D Coval, and Ľuboš Pástor, 2005, Judging Fund Managers by the Company They Keep, *The Journal of Finance* 60, 1057–1096.

Conejo, A. J., M. A. Plazas, R. Espinola, and A. B. Molina, 2005, Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models, *IEEE Transactions on Power Systems* 20, 1035–1042.

Corsi, Fulvio, 2009, A Simple Approximate Long-Memory Model of Realized Volatility, *Journal of Financial Econometrics* 7, 174–196.

Coval, Joshua D, and Tobias J Moskowitz, 1999, Home Bias at Home: Local Equity Preference in Domestic Portfolios, *The Journal of Finance* 54, 2045–2073.

Coval, Joshua D, and Tobias J Moskowitz, 2001, The Geography of Investment: Informed Trading and Asset Prices, *The Journal of Political Economy* 109, 811–841.

Del Negro, Marco, Raiden B. Hasegawa, and Frank Schorfheide, 2016, Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance, *Journal of Econometrics* 391–405.

Diebold, Francis X., Todd A. Gunther, and Anthony S. Tay, 1998, Evaluating Density Forecasts with Applications to Financial Risk Management, *International Economic Review* 39, 863–883.

Driscoll, John C., and Aart C. Kraay, 1998, Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data, *Review of Economics and Statistics* 80, 549–560.

Dybvig, Philip H., and Stephen A. Ross, 1985, Differential Information and Performance Measurement Using a Security Market Line, *The Journal of Finance* 40, 383–399.

Elton, Edwin J, Martin J Gruber, and Christopher R Blake, 1996, The Persistence of Risk-Adjusted Mutual Fund Performance, *The Journal of Business* 69, 133–157.

Fama, Eugene F, and Kenneth R French, 1993, Common Risk Factors in the Returns on Stocks and Bonds, *Journal of Financial Economics* 33, 3–56.

Fang, Lily H., Joel Peress, and Lu Zheng, 2014, Does Media Coverage of Stocks Affect Mutual Funds' Trading and Performance?, *Review of Financial Studies* 27, 3441–3466.

Fliedner, Gene, 1999, An Investigation of Aggregate Variable Time Series Forecast Strategies with Specific Subaggregate Time Series Statistical Correlation, *Computers & Operations Research* 26, 1133–1149.

Gençay, Ramazan, Faruk Selçuk, and Brandon J Whitcher, 2001, *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics* (Elsevier).

Gençay, Ramazan, Daniele Signori, Yi Xue, Xiao Yu, and Keyi Zhang, 2015, Economic Links and Credit Spreads, *Journal of Banking & Finance* 55, 157–169.

Grinblatt, Mark, and Sheridan Titman, 1992, The Persistence of Mutual Fund Performance, *The Journal of Finance* 47, 1977–1984.

Grinblatt, Mark, Sheridan Titman, and Russ Wermers, 1995, Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior, *American Economic Review* 85, 1088–1105.

Gupta-Mukherjee, Swasti, and Ankur Pareek, 2014, Limited Attention and Portfolio Choice: The Impact of Attention Allocation on Mutual Fund Performance, in *AFA 2013 San Diego Meetings Paper*.

Hansen, Lars Peter, 1982, Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* 50, 1029–1054.

Hausman, Jerry A., and William E. Taylor, 1981, Panel Data and Unobservable Individual Effects, *Econometrica* 49, 1377–1398.

Hayashi, Fumio, 2000, *Econometrics* (Princeton University Press, Princeton, N.J).

Hendershott, Terrence, Dmitry Livdan, and Norman Schürhoff, 2015, Are Institutions Informed About News?, *Journal of Financial Economics* 117, 249–287.

Hertzel, M, Z Li, M Officer, and K Rodgers, 2008, Inter-Firm Linkages and the Wealth Effects of Financial Distress Along the Supply Chain, *Journal of Financial Economics* 87, 374–387.

Huberman, Gur, and Tomer Regev, 2001, Contagious Speculation and a Cure for Cancer: A Nonevent That Made Stock Prices Soar, *The Journal of Finance* 56, 387–396.

Huij, Joop, and Marno Verbeek, 2007, Cross-Sectional Learning and Short-Run Persistence in Mutual Fund Performance, *Journal of Banking & Finance* 31, 973–997.

Hyndman, Rob J., Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang, 2011, Optimal Combination Forecasts for Hierarchical Time Series, *Computational Statistics & Data Analysis* 55, 2579–2589.

Hyndman, Rob J., and Yeasmin Khandakar, 2008, Automatic Time Series Forecasting: The Forecast Package for R, *Journal of Statistical Software* 27, 1–22.

Hyndman, Rob J., Alan J. Lee, and Earo Wang, 2016, Fast Computation of Reconciled Forecasts for Hierarchical and Grouped Time Series, *Computational Statistics & Data Analysis* 97, 16–32.

Investment Company Institute, 2014, Investment Company Fact Book: A Review of Trends and Activities in the Investment Company Industry, Technical report, Investment Company Institute, Washington, D.C.

Kacperczyk, Marcin, and Amit Seru, 2007, Fund Manager Use of Public Information: New Evidence on Managerial Skills, *The Journal of Finance* 62, 485–528.

Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng, 2005, On the Industry Concentration of Actively Managed Equity Mutual Funds, *The Journal of Finance* 60, 1983–2011.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp, 2009, Rational Attention Allocation Over the Business Cycle, Technical report, National Bureau of Economic Research.

Kahneman, Daniel, 1973, *Attention and effort* (Citeseer).

Kleibergen, Frank, and Richard Paap, 2006, Generalized Reduced Rank Tests Using the Singular Value Decomposition, *Journal of Econometrics* 133, 97–126.

Kleibergen, Frank, and Mark E. Schaffer, 2015, RANKTEST: Stata Module to Test the Rank of a Matrix Using The Kleibergen-Paap Rk Statistic.

Kohn, Robert, 1982, When Is an Aggregate of a Time Series Efficiently Forecast by Its Past?, *Journal of Econometrics* 18, 337–349.

Lesmond, David A., Michael J. Schill, and Chunsheng Zhou, 2004, The Illusory Nature of Momentum Profits, *Journal of Financial Economics* 71, 349–380.

Lettau, Martin, and Sydney Ludvigson, 2001, Consumption, Aggregate Wealth, and Expected Stock Returns, *The Journal of Finance* 56, 815–849.

Lintner, John, 1965, The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, *Review of Economics and Statistics* 47, 13–37.

Lückoff, Peter, 2011, *Mutual Fund Performance and Performance Persistence: The Impact of Fund Flows and Manager Changes* (Springer).

Mele, Antonio, 2007, Asymmetric Stock Market Volatility and the Cyclical Behavior of Expected Returns, *Journal of Financial Economics* 86, 446–478.

Merton, Robert C., 1974, On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, *The Journal of Finance* 29, 449.

Nanda, Vikram, Z. Jay Wang, and Lu Zheng, 2004, Family Values and the Star Phenomenon: Strategies of Mutual Fund Families, *Review of Financial Studies* 17, 667–698.

Newey, Whitney K., and Kenneth D. West, 1987, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55, 703–708.

Orphanides, Athanasios, and Simon van Norden, 2005, The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time, *Journal of Money, Credit and Banking* 37, 583–601.

Pareek, Ankur, 2012, Information Networks: Implications for Mutual Fund Trading Behavior and Stock Returns, in *AFA 2010 Atlanta Meetings Paper*.

Paye, Bradley S., 2012, 'Déjà Vol': Predictive Regressions for Aggregate Stock Market Volatility Using Macroeconomic Variables, *Journal of Financial Economics* 106, 527–546.

Rapach, David E., Jack K. Strauss, and Guofu Zhou, 2010, Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy, *The Review of Financial Studies* 23, 821–862.

Schwert, G. William, 1989, Why Does Stock Market Volatility Change Over Time?, *The Journal of Finance* 44, 1115.

Sharpe, William F, 1964, Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk, *The Journal of Finance* 19, 425–442.

Simon, Herbert A, 1971, Designing Organizations for an Information-Rich World, in Martin Greenberger, ed., *Computers, Communications, and the Public Interest*, 40–41 (Johns Hopkins University Press).

Staiger, Douglas, and James H. Stock, 1997, Instrumental Variables Regression with Weak Instruments, *Econometrica* 65, 557–586.

Stock, James H., and Mark W. Watson, 2002, Macroeconomic Forecasting Using Diffusion Indexes, *Journal of Business & Economic Statistics* 20, 147–162.

Stock, James H., and Mark W. Watson, 2003, Forecasting Output and Inflation: The Role of Asset Prices, *Journal of Economic Literature* 41, 788–829.

Stock, James H., and Mark W. Watson, 2004, Combination Forecasts of Output Growth in a Seven-Country Data Set, *Journal of Forecasting* 23, 405–430.

Ter Horst, Jenke, and Marno Verbeek, 2000, Estimating Short-Run Persistence in Mutual Fund Performance, *Review of Economics and Statistics* 82, 646–655.

Tiao, G. C., and Irwin Guttman, 1980, Forecasting Contemporal Aggregates of Multiple Time Series, *Journal of Econometrics* 12, 219–230.

Veldkamp, Laura L., 2006, Media Frenzies in Markets for Financial Information, *The American Economic Review* 577–601.

Wermers, Russ, 2000, Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses, *The Journal of Finance* 55, 1655–1703.

Yuan, Yu, 2015, Market-Wide Attention, Trading, and Stock Returns, *Journal of Financial Economics* 116, 548–564.

Zellner, Arnold, and Justin Tobias, 2000, A Note on Aggregation, Disaggregation and Forecasting Performance, *Journal of Forecasting* 19, 457–465.