International Review of Economics and Finance 71 (2021) 299–315

Contents lists available at ScienceDirect

International Review of Economics and Finance

journal homepage: www.elsevier.com/locate/iref

What can cluster analysis offer in investing? - Measuring structural changes in the investment universe

Min Kyu Sim [a,*], Shijie Deng [b], Xiaoming Huo [b]

[a] Seoul National University of Science and Technology, Department of Industrial Engineering, 232 Gongneung-ro, Nowon-gu, Seoul, 01811, Republic of Korea

[b] Georgia Institute of Technology, School of Industrial and Systems Engineering, 765 Ferst Drive, NW, Atlanta, GA, 30332, USA

ARTICLE INFO

JEL classification:
JEL
C10
C58
D85
G11
G12

Keywords:
Cluster analysis
Investment opportunity set
Basis assets
Asset pricing model
Factor model

ABSTRACT

The return on assets of the investment universe tends to form a cluster structure. This study quantifies this strength of the clustering tendency as a single econometric measure, referred to as modularity. Through an empirical study of the US equity market, we demonstrate that the strength of the clustering tendency changes over time with market fluctuations. That is, normal markets tend to have a clear cluster structure (high modularity), while stressed markets tend to have a blurry cluster structure (low modularity). Modularity assesses the quality of an investment opportunity set in terms of potential diversification benefits. Modularity is an important pricing variable in the cross-sectional returns of US stocks. From 1992 to 2015, the average return of the stocks with the lowest sensitivity to modularity (low modularity beta) exceeds that of the stocks with the highest sensitivity (high modularity beta) by approximately 10.49% annually, adjusted for the Fama-French five-factor exposures. The inclusion of modularity as an asset pricing factor, therefore, expands the investment opportunity set for factor-based investors.

1. Introduction

The presence of more investable assets is likely to expand the investment opportunity set, and thus improve investors' utility. However, an important pre-condition is a moderate level of co-movement tendency, which effectively reduces the overall portfolio risk by allowing diversification effects. Since the scope of financial markets has expanded with the advent of innovative financial securities and various derivative products, it is crucial for investors to comprehend the structural co-movement tendency of the investment universe and assess the investment opportunity set.

This study provides an investment framework that visualizes and quantifies the co-movement structure. Through our empirical studies, we demonstrate that this framework can 1) assess the quality of an investment opportunity set with regard to the potential benefits from diversification, 2) generate an asset pricing factor, and 3) expand the investment opportunity set for factor-based investors.

We focus on and combine two well-known structural properties of the investment universe. First, the co-movement structure of financial assets is time-varying. Billio et al. (2012) and Diebold and Yilmaz (2014) graphically depict increasing level of associations between financial institutions during the global financial crisis from 2007 to 2008. Further, Buraschi et al. (2010) and Sandoval Jr. and Franca (2012) note rises in correlations in a similar period. By adopting the perspective of the clustering property, we develop insight into time variability in terms of the level of associations among financial assets.

* Corresponding author.
E-mail address: mksim@seoultech.ac.kr (M.K. Sim).

https://doi.org/10.1016/j.iref.2020.09.004
Received 5 November 2018; Received in revised form 12 July 2020; Accepted 4 September 2020
Available online 18 September 2020

Second, financial assets generally form cluster structures. The clustering property, by which entities with similar characteristics tend to form a subgroup is one of the most evident structural properties in a network of financial assets. Materassi and Innocenti (2009) provide empirical evidence that major US stocks can be arranged into a tree-shaped diagram where highly correlated stocks are grouped by the branches of the tree. Jang et al. (2011) apply the minimum spanning tree algorithm to depict structural changes in foreign exchange markets.

The cluster analysis we perform involves grouping a set of financial assets into subgroups according to their co-movement tendency. This approach is often called *identification of community structure*, and it is clearly distinguished from data clustering algorithms typically represented by K-means and K-nearest neighbors.

We aim to grasp the entire network structure formed by the set of financial assets and are again distinguished from dimension reduction techniques. Dimension reduction techniques aim to summarize network structures with a few variables, but our approach, based on clustering analysis, creates a single variable that merely measures the clustering tendency of the entire network instead of dealing with individual subgroups. This distinction is indeed the central technical motivation of our study in that most dimension reduction techniques for summarizing original variables into a few numbers of variables - whether main principal component vectors in PCA or common drivers of graphical models - still leave unexplained parts that form a clustered structure. For example, Chandrasekaran et al. (2012) conduct an empirical test with their hidden Gaussian graphical model and demonstrate that the clustering tendency across US stocks is still persistent after some common hidden drivers are identified and their influences are removed. If a dimension reduction method goes further in order to tackle the issue of the clustering tendency in the unexplained parts, it has to spend an additional dimension for each clustered subgroup. This diminishes the beauty of succinct modeling which ought to be achieved by dimension reduction.

Once clustered subgroups of assets are identified through cluster analysis, we classify all the pair-wise correlations of the assets into two sets. The first set collects the correlations between assets in the same subgroups, and the second set collects the correlations between assets in the different subgroups. The difference between the average correlations of each set, termed *modularity*, demonstrates the strength of the cluster structure and can be measured using time series data for asset returns (Section 2). The obtained modularity is related to market fluctuation and the quality of the investment opportunity set with respect to the potential diversification benefit (Section 3). The measure can serve as a priced state variable; in other words, how an individual asset is related to modularity explains the average return (Section 4). Modularity as an asset pricing factor can expand the investment opportunity set for factor-based investments (Section 5).

This study contributes in three ways. First, we extend studies of the associations of financial assets (Pollet and Wilson (2010), Jang et al. (2011), Billio et al. (2012), and Diebold and Yilmaz (2014)) by applying the perspective of clustered networks and proposing a quantitative model at the individual stock or portfolio level. Second, given that the previous studies (Materassi and Innocenti (2009) and Chandrasekaran et al. (2012)) have found cluster analysis to be useful for visualizing the structure of financial assets, we further expand the applicability of these visualizations to a framework for asset management. Third, Ahn et al. (2009) discuss grouping of stocks and investment scenarios with grouped basis assets. We add time varying aspects of grouping and introduce an investment framework in which the time variability of the network is incorporated into the set of basis assets composed of established pricing factors (Fama and French (1992, 1993, 2015)). Lastly, we propose an additional factor to the studies on asset pricing factors. Our proposed asset pricing factor is distinguishably based on the interaction between an individual security and the market structure, unlike other asset pricing factors based on the characteristics of individual firms such as stock price and financial ratios (Fama and French (1992, 1993, 2015); Carhart (1997); Jegadeesh and Titman (1993)).

2. Construction of the connectedness measures

The research strand on the associations among financial assets, such as Billio et al. (2012) and Diebold and Yilmaz (2014), adopts the graph and network perspectives to assess the overall market structure and its changes over time. We follow a bottom-up approach to construct connectedness measures that ultimately model the clustering tendency of asset returns. Although our framework and definition of connectedness can be applied to general financial assets, we confine our attention to modeling stock returns.

We initiate the construction process by measuring the *connectedness of two stocks*. Let $C(i,j)$ denote the Pearson's pair-wise correlation between the two stocks:

$$C(i,j) := \rho_{i,j} = \frac{Cov(r_i, r_j)}{s.d.(r_i)s.d.(r_j)}, \tag{1}$$

where r_i and r_j denote the returns of stock i and j, respectively. This measure is extended to define the *connectedness between two groups of stocks* by considering all possible combinations of stocks in each group. Namely,

$$C(A,B) := avg(\{C(i,j)|(i,j) \in (A,B), i \neq j\}), \tag{2}$$

where A and B are two groups of stocks, and the operator,[1] $avg(\cdot)$, calculates the average value of the elements in the set. Note that

[1] Taking the average of the correlation elements is not a mathematically straightforward task. One may simply think of taking the arithmetic average of all correlations under consideration; however, Zimmerman et al. (2003) show that this may result in a biased estimator. In a geometrical sense, Pearson's pair-wise correlation is a cosine value of the angle formed by two vectors, and is thus not additive. As a result, the variance stabilization methods of Fisher's transformation (Fisher (1915) and Fisher (1924)) should be used to estimate the average value. We accordingly define an averaging operator as $avg(r_1, r_2, \ldots, r_n) := (exp(2\bar{z}) - 1)/(exp(2\bar{z}) + 1)$, where $r_1, r_2, \ldots, r_n$ are the sample correlation elements and $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} 0.5\, log\left(\frac{1+r_i}{1-r_i}\right)$

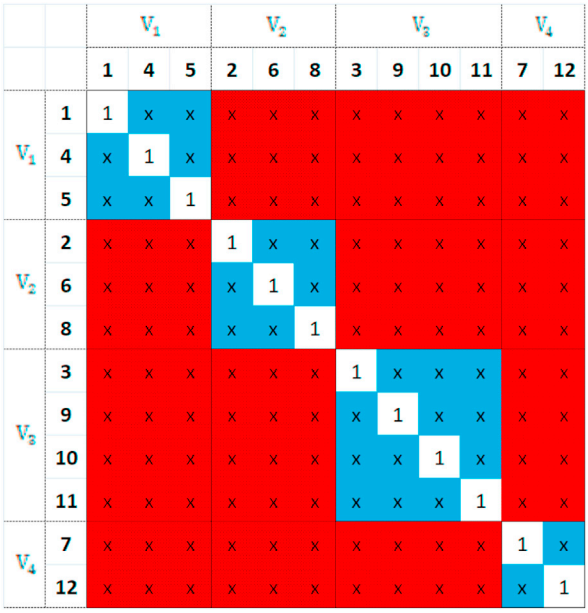

Fig. 1. Construction of the connectedness measures. After the variables are reordered according to partition P, INSC(P) is the average of the elements on the diagonal blocks (blue solid background) and ITSC(P) is the average of the elements on the off-diagonal block (red dotted background). If partition P accurately describes the cluster structure of the universe, then INSC(P) must be higher than ITSC(P). It follows that high MOD(P) implies well-clustered structure with respect to partition P.

groups A and B are allowed to have overlapping elements (or can be even identical) due to the condition, $i \neq j$, which excludes trivial self-correlations for the overlapping stocks. We note that the *connectedness of the two groups* can be used to define the *total system connectedness* of an entire stock market. Namely, let V be a group containing all the stocks in a market, then $C(V, V)$ is the average of all the pair-wise correlations among all stocks. Pollet and Wilson (2010) demonstrate that the average correlation of all constituent stocks for the S&P500 Index can predict the future quarterly returns of the index.[2]

Now that our measures cover groups of stocks, we introduce the notions of *partition* and *cluster* to analyze multi-group structures in the investment universe. Let V denote the set of stocks under consideration. Partition P of set V is a grouping of the set elements such that 1) each subgroup in P is not empty, 2) the subgroups are mutually exclusive, and 3) the union of all subgroups is equal to set V. In other words, $P = \{V_1, V_2, \ldots, V_k\}$ where $V_i \neq \varnothing$ for all i, $V_i \cap V_j = \varnothing$ for all $i \neq j$, and $V = \cup_{c=1}^{k} V_c$. *Cluster analysis or clustering* on a set of correlated variables is the task of finding the best partition, P, for set V such that the correlations among the stocks within each subgroup are higher than the correlations among the stocks that belong to different subgroups.

Assuming a suitable partition P is found (the clustering method used in our study is discussed at the end of this section), the remaining connectedness measures the investment universe V are defined with respect to fixed partition P. *Inner-sector connectedness (INSC)* is defined as the average of all pair-wise correlations within the subgroups in the partition $P = \{V_1, V_2, \ldots, V_k\}$:

$$INSC(P) : = avg\Big(\bigcup_{c=1}^{k} \{C(i,j) | (i,j) \in (V_c, V_c), i \neq j\} \Big). \tag{3}$$

Similarly, *inter-sector connectedness (ITSC)* is defined as the average of all correlations across the subgroups in partition $P = \{V_1, V_2, \ldots, V_k\}$:

$$ITSC(P) : = avg\Big(\bigcup_{c_1=1}^{k-1} \bigcup_{c_2=c_1+1}^{k} \{C(i,j) | (i,j) \in (V_{c_1}, V_{c_2})\} \Big) \tag{4}$$

Fig. 1 illustrates a correlation matrix of 12 random variables with partition $P = \{(1,4,5),(2,6,8),(3,9,10,11),(7,12)\}$. After the variables for the correlation matrix are reordered according to partition P, INSC(P) is the average value of the elements on the diagonal blocks (blue solid background) and ITSC(P) is the average value of the elements on the off-diagonal blocks (red dotted background).

As long as partition P prominently describes the cluster structure, INSC(P) is expected to be much higher than ITSC(P) meaning that

[2] We note that the total system connectedness is the weighted average of INSC(P) and ITSC(P), which are to be defined shortly. That is, $TSC(P) = \frac{a \cdot INSC(P) + b \cdot ITSC(P)}{a+b}$, where a is the number of correlation elements used to generate INSC(P), and b is the number of correlation elements used to generate ITSC(P). More specifically, $a = \sum_{i=1}^{k} |V_i|(|V_i| - 1)/2$ and $b = \sum_{c_1=1}^{k-1} \sum_{c_2=c_1+1}^{k} |V_{c_1}||V_{c_2}|$, where $P = \{V_1, \ldots, V_k\}$ and $|V_i|$ denotes the number of element in subgroup V_i.

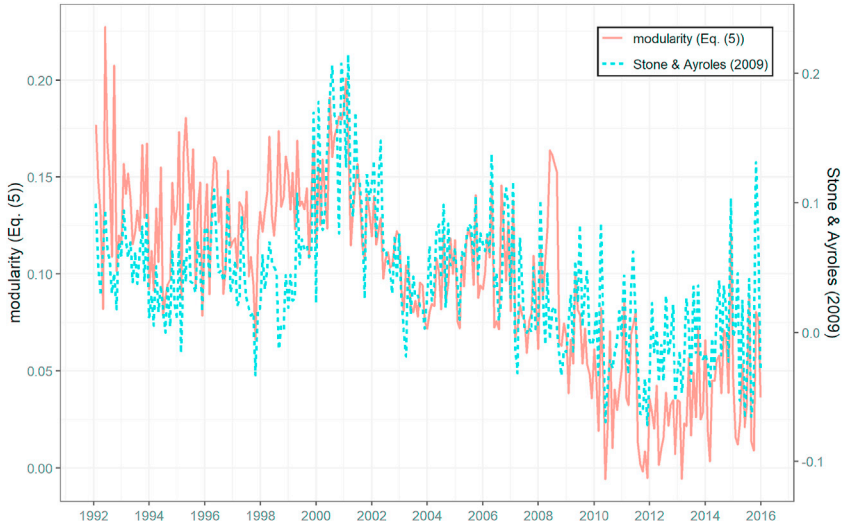

Fig. 2. Fluctuation patterns of monthly modularity vs. the measure of Stone and Ayroles (2009). The relative innovations of the two measures are similar across the time.

the stocks should have a much higher correlation within their subgroups than against the stocks of other subgroups. In other words, high INSC(P) in conjunction with low ITSC(P) implies that universe V is well clustered with respect to the partition P, while low INSC(P) combined with a high ITSC(P) implies the contrary. MOD(P), defined as the difference between INSC(P) and ITSC(P), thus concisely quantifies the strength of the clustering tendency as a single measure:

$$MOD(P) := INSC(P) - ITSC(P) \tag{5}$$

Identifying the most appropriate cluster structure is obviously a crucial prerequisite for constructing meaningful connectedness measures. In a single stock market, classifying stocks into subgroups by using traditional Standard Industrial Classification (SIC) codes is a widely accepted and plausible practice. However, technical approaches based on correlations between stock returns are more suitable for investors to assess the investment opportunity set based on possible diversification benefits.[3]

For the empirical tests throughout the rest of this study, we adopt the modulated modularity clustering (MMC) method proposed by Stone and Ayroles (2009) for two main reasons. First, unlike most other clustering algorithms, the MMC algorithm does not require user's input of the number of subgroups in advance. The MMC algorithm takes only input regarding the relative fineness and coarseness of the cluster solution, and the number and size of the subgroups are determined by the strength of its subgroups' clustering tendency.[4] Second, the embedded spectral decomposition process of the MMC algorithm efficiently handles high-dimensional data.

Since the clustering algorithm plays such an important role in our empirical tests, we briefly review the graph theory and MMC algorithm in the remaining part of this section. In a graph, partitioning[5] is a task that serves to identify disjoint subsets of nodes in which each subset (called a *cell*) has nodes that are close to each other, and the nodes are not close to each other across the different cells.

In general, heuristic partitioning algorithms utilize a top-down or a bottom-up approach with some stopping criteria. However, the line of studies including Newman and Girvan (2004), Newman (2006), and Stone and Ayroles (2009) contribute by setting a partitioning problem as a combinatorial problem of a single objective function where solutions can be obtained through analytic procedures. A seminal work in this field Newman and Girvan (2004) defines modularity as a single quantity for clustering an unweighted 0–1 graph, and Newman (2006) provides a succinct description: "The modularity is, up to multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random.'' (p. 8578) Newman (2006) proceeds to develop an algorithm that maximizes the modularity defined above. For a weighted graph, such as the network modelled by Pearson's correlation matrix, Stone and Ayroles (2009) define the equivalent notion of modularity for weighted graphs as the sum of the edge weights within groups minus the expected sum of the edges weights in an equivalent network with edges placed at random. Stone and Ayroles (2009) provide a spectral decomposition algorithm that maximizes their notion of modularity.

The modularities defined by Newman and Girvan (2004) and Stone and Ayroles (2009) not only serve as an objective function, but

[3] As expected, our empirical tests with SIC grouping led to similar but weaker results.

[4] We used the fineness controlling parameter σ in Stone and Ayroles (2009) set to be 0.45, resulting the number of groups for 60 stocks to be 7–15 groups in the test periods. Other tried parameter values for σ, ranging from 0.2 to 0.6, lead to similar test results with different numbers of groups, but we believe that grouping 60 stocks into 7–15 groups is in accordance with general grouping practice for investing. How many groups are appropriate to split the universe of stocks in an economy is always subject to debate. The important feature of MMC grouping is, however, the flexibility of the number of groups generated depending on the network itself with same controlling parameter value.

[5] Interchangeable terms include *clustering* and *community structure identification* depending on the fields of study.

Table 1
Full list of representative stocks grouped by MMC algorithm. The 60 stocks are selected from the top of the Fortune 500 list published in 2015, which ranks US companies by operating revenue during 2014. The identified structure is similar but not identical to standard industry classification.

Cell	Ticker	Company Name	Operating Revenue in 2014 (MM)	SIC Code	Industry
1	PG	Procter & Gamble	78,756	2841	Soap and Other Detergents, except Specialty Cleaners
1	JNJ	Johnson & Johnson	70,074	2834	Pharmaceutical Preparations
1	PEP	PepsiCo	63,056	2086	Bottled & Canned Soft Drinks & Carbonated Waters
1	PFE	Pfizer	48,851	2834	Pharmaceutical Preparations
1	KO	Coca-Cola	44,294	2086	Bottled & Canned Soft Drinks & Carbonated Waters
2	BRK	Berkshire Hathaway	210,821	6331	Fire, Marine & Casualty Insurance
2	F	Ford Motor	149,558	3711	Motor Vehicles & Passenger Car Bodies
2	CMCSA	Comcast	74,510	4840	Cable and Other Pay Television Services
2	UPS	UPS	58,363	4513	Air Courier Services
2	DOW	Dow Chemical	48,778	2821	Plastic Materials, Synth Resins & Nonvulcan Elastomers
2	FDX	FedEx	47,453	4513	Air Courier Services
2	TSN	Tyson Foods	41,373	2015	Poultry Slaughtering and Processing
2	JCI	Johnson Controls	40,204	2531	Public Bldg & Related Furniture
3	XOM	Exxon Mobil	246,204	2911	Petroleum Refining
3	CVX	Chevron	131,118	2911	Petroleum Refining
3	PSX	Phillips 66	87,169	1382	Oil & Gas Field Exploration Services
3	VLO	Valero Energy	81,824	2911	Petroleum Refining
3	MPC	Marathon Petroleum	64,566	2911	Petroleum Refining
4	GE	General Electric	140,389	3600	Electronic & Other Electrical Equipment (No Computer Equip)
4	JPM	J.P. Morgan Chase	101,006	6021	National Commercial Banks
4	BAC	Bank of America Corp.	93,056	6021	National Commercial Banks
4	WFC	Wells Fargo	90,033	6021	National Commercial Banks
4	C	Citigroup	88,275	6021	National Commercial Banks
4	MET	MetLife	69,951	6311	Life Insurance
4	AIG	AIG	58,327	6331	Fire, Marine & Casualty Insurance
4	PRU	Prudential Financial	57,119	6311	Life Insurance
5	MCK	McKesson	181,241	5122	Wholesale-Drugs, Proprietaries & Druggists' Sundries
5	ABC	AmerisourceBergen	135,962	5122	Wholesale-Drugs, Proprietaries & Druggists' Sundries
5	CAH	Cardinal Health	102,531	5122	Wholesale-Drugs, Proprietaries & Druggists' Sundries
5	ESRX	Express Scripts Holding	101,752	8093	Services-Specialty Outpatient Facilities, NEC
Cell	Ticker	Company Name	Operating Revenue in 2014 (MM)	SIC Code	Industry
6	HD	Home Depot	88,519	5211	Retail-Lumber & Other Building Materials Dealers
6	TGT	Target	73,785	5331	Retail-Variety Stores
6	LOW	Lowe???s	59,074	5211	Retail-Lumber & Other Building Materials Dealers
7	BA	Boeing	96,114	3721	Aircraft
7	ADM	Archer Daniels Midland	67,702	2070	Fats & Oils
7	UTX	United Technologies	61,047	3724	Aircraft Engines & Engine Parts
7	DIS	Disney	52,465	4841	Cable & Other Pay Television Services
7	CAT	Caterpillar	47,011	3531	Construction Machinery & Equip
8	HPQ	HP	103,355	3571	Electronic Computers
8	MSFT	Microsoft	93,580	7370	Services-Computer Programming, Data Processing, Etc.
8	IBM	IBM	82,461	3570	Computer & office Equipment
8	INTC	Intel	55,355	3679	Electronic Components, NEC
8	CSCO	Cisco Systems	49,161	3674	Semiconductors & Related Devices
8	IM	Ingram Micro	43,026	5045	Wholesale-Computers & Peripheral Equipment & Software
9	UNH	UnitedHealth Group	157,107	6324	Hospital & Medical Service Plans
9	ANTM	Anthem	79,157	6324	Hospital & Medical Service Plans
9	AET	Aetna	60,337	6324	Hospital & Medical Service Plans
9	HUM	Humana	54,289	6324	Hospital & Medical Service Plans
10	WMT	Walmart	482,130	5331	Retail-Variety Stores
10	CVS	CVS Health	153,290	5912	Retail-Drug Stores and Proprietary Stores
10	COST	Costco	116,199	5331	Retail-Variety Stores
10	KR	Kroger	109,830	5411	Retail-Grocery Stores
10	WBA	Walgreens Boots Alliance	103,444	5912	Retail-Drug Stores and Proprietary Stores
10	SYY	Sysco	48,681	5140	Wholesale-Groceries & Related Products
10	LMT	Lockheed Martin	46,132	3760	Guided Missiles & Space Vehicles & Parts
11	AAPL	Apple	233,715	3571	Electronic Computers
11	AMZN	Amazon.com	107,006	7370	Services-Computer Programming, Data Processing, Etc.
11	GOOG	Alphabet	74,989	7375	Information Retrieval Services
12	T	AT&T	146,801	4813	Telephone Communications (No Radiotelephone)
12	VZ	Verizon	131,620	4813	Telephone Communications (No Radiotelephone)

they can also measure the fitness of a network with respect to a certain fixed partition (i.e., clustering strength). Emphasizing the aspect of clustering strength, we propose a measure of modularity as the average weight of the edges of inner groups minus the average weight of the edges between groups. Our measure and the modularity measure by Stone and Ayroles (2009) share motivation, and the relative innovations are close. Fig. 2 presents the closeness of the two measures. It is noticeable that the general fluctuating patterns of two are

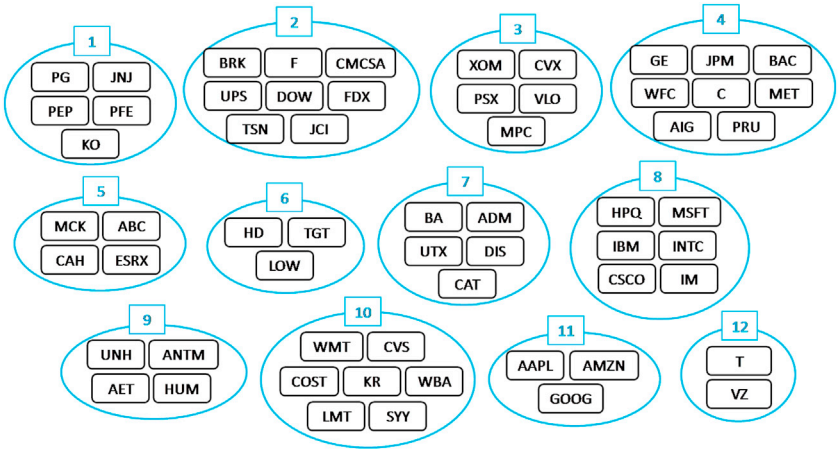

Fig. 3. Cluster structure of the 60 major companies using the MMC algorithm (2005–2014). Clustering is conducted to combine stocks with high correlations in the same subgroup and separate stocks with low correlations into different subgroups. The full list with company description is presented in Table 1. This grouping serves as a fixed point for analysis in Section 3.

very similar across the time.

While the measure by Stone and Ayroles (2009) enjoys the benefits of the analytic algorithm, our measure is easier to use and understand because we apply the notion of the difference between two average values: the inner-sector connectedness (INSC) and inter-sector connectedness (ITSC). By no means do we claim that our modularity is always superior to the measure of Stone and Ayroles (2009). In fact, how the general notions of modularity should be defined and measured is subject to the nature of individual networks under the consideration. We claim that our measure is a significant and intuitive means of providing an investment framework regarding the structure of financial markets.

3. Assessing the set of investment opportunity - marketwide empirical evidence

This section empirically studies the clustering tendency of US stocks in light of the defined connectedness measure, and it assesses the time-varying quality of the investment opportunity set with respect to the potential benefits of portfolio diversification. To assess the time-varying and qualitative features, it is necessary to set up a time-invariant partition before investigating data for different time periods. Thus, Section 3.1 uses the long-term (10-year) sample correlation matrix to generate a cluster solution that acts as the time-invariant partition when analyzing subperiods. Using the time-invariant cluster solution, Section 3.2 applies the defined connectedness measures to the selected subperiods that displayed dramatic market fluctuation.

This section takes 60 major companies (Table 1 presents the list) as a representative set and uses their daily stock return data over the 10-year period from 2005 to 2014. The stocks are selected from the top of the Fortune 500 list announced in 2015, which ranks companies by operating revenue in 2014. We believe that the top of the Fortune 500 list should serve as a representative set for the domestic economy. Although the top rankers in market capitalization can also be considered, they tend to be heavily focused on a few capital-oriented industries, such as the petroleum refining industry and financial services. The return data were obtained from the Center for Research in Security Prices (CRSP) database provided by Wharton Research Data Services (WRDS). Stocks must be common shares (WRDS share code 10 and 11), and stocks with missing daily returns information at any point in the decade are excluded from the analysis.[6]

3.1. Cluster of asset returns in the long term

By using the daily returns of the 60 stocks from January 1, 2005 to December 31, 2014, a historical correlation matrix is prepared. Applying the MMC algorithm to the historical correlation matrix identifies the cluster structure presented in Fig. 3. The identified cluster structure is similar, but not identical, to those categorized using SIC codes. This partition solution alone can serve as a useful decision-making ground, if an investor's practice of grouping his/her investment universe is more concerned with the technical grouping rather than traditional industry grouping.

Fig. 4, as a close observation of Fig. 3, presents selected subgroups (the 4th, 9th, and 11th subgroups in Table 1) with *connectedness between two groups* as defined in (2). The connectedness measures within the same subgroups (on the blue solid lines; $C(V_4, V_4) = 0.66$, $C(V_9, V_9) = 0.69$, $C(V_{11}, V_{11}) = 0.45$) are higher than the connectedness measures across the different subgroups (on the red dotted

[6] We admit that the analysis of this section has the bias issues of lookback (stocks are selected at the end of the analysis period) and survivorship (stocks with missing data are excluded from the analysis). This section provides a qualitative understanding of the defined measures, and empirical tests in the following sections are free of such biases.

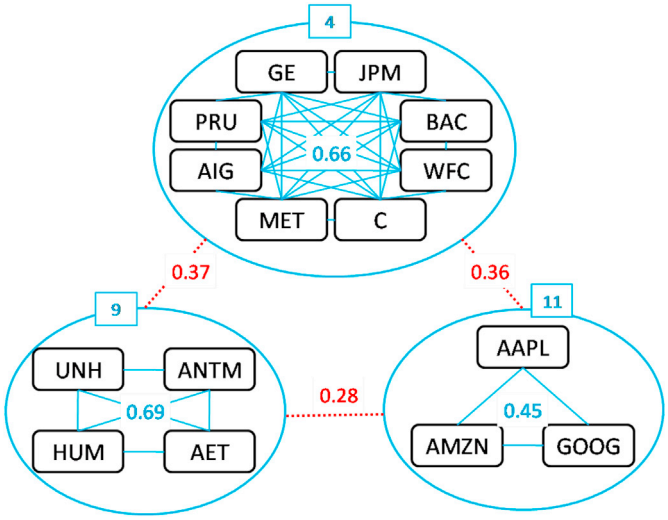

Fig. 4. A close-up view of three selected subgroups from Fig. 3. The blue solid lines indicate correlations within each subgroup and the red dotted lines indicate correlations across different subgroups. The connectedness measures within subgroups are higher than those across different subgroups, indicating that the structure is well clustered with respect to the partition presented in Fig. 3 and Table 1.

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.555	0.420	0.461	0.399	0.434	0.430	0.476	0.425	0.377	0.429	0.329	0.511
2	0.420	0.477	0.457	0.472	0.376	0.476	0.506	0.446	0.351	0.390	0.364	0.451
3	0.461	0.457	0.768	0.448	0.409	0.425	0.550	0.464	0.373	0.384	0.390	0.486
4	0.399	0.472	0.448	0.658	0.364	0.476	0.483	0.431	0.368	0.364	0.363	0.450
5	0.434	0.376	0.409	0.364	0.542	0.379	0.419	0.380	0.427	0.378	0.302	0.393
6	0.430	0.476	0.425	0.476	0.379	0.687	0.489	0.448	0.350	0.464	0.401	0.484
7	0.476	0.506	0.550	0.483	0.419	0.489	0.568	0.493	0.384	0.431	0.409	0.494
8	0.425	0.446	0.464	0.431	0.380	0.448	0.493	0.519	0.329	0.385	0.410	0.468
9	0.377	0.351	0.373	0.368	0.427	0.350	0.384	0.329	0.688	0.327	0.281	0.342
10	0.429	0.390	0.384	0.364	0.378	0.464	0.431	0.385	0.327	0.422	0.322	0.424
11	0.329	0.364	0.390	0.363	0.302	0.401	0.409	0.410	0.281	0.322	0.450	0.382
12	0.511	0.451	0.486	0.450	0.393	0.484	0.494	0.468	0.342	0.424	0.382	0.749

Fig. 5. Matrix representation for the correlation structure of Fig. 3. The contrast between the on- and off-diagonal elements is noticeable, indicating a prominent cluster structure and a high value of MOD(P).

lines; $C(V_4, V_9) = 0.37$, $C(V_4, V_{11}) = 0.36$, $C(V_9, V_{11}) = 0.28$). This relative difference between the same and different subgroups confirms that the clustering is effectively conducted.

Fig. 5 displays all of the connectedness measures between groups in a matrix form, $[C(V_i, V_j)]_{1\le i,j\le 12}$. This matrix representation provides a simple view of the entire structure with the connectedness measures. It summarizes the original 60×60 correlation matrix into a 12×12 matrix that shows clear contrasts between the diagonal and off-diagonal elements. The following are equivalent: 1) a high value of MOD(P), 2) a prominent cluster structure, and 3) a noticeable difference between the on- and off-diagonals in the matrix representation of Fig. 5.

3.2. Subperiods of bull/bear markets

The representative set of investable stocks above displays a cluster structure in the long term. The obvious next question is whether the cluster structure in the long run is persistent throughout its subperiods. This subsection selects four subperiods with drastic market movements, of which two are bullish and the other two are bearish, and observes the clustering tendency of each subperiod. To consistently compare the subperiods, the partition structure obtained from the long-term analysis (Fig. 3) is maintained for all four subperiods.

Fig. 6 displays the matrices of subgroup connectedness in each subperiod, and Fig. 7 presents connectedness diagrams for the previously selected 4th, 9th, and 11th subgroups. In both Figs. 6 and 7, the two top diagrams correspond to bullish subperiods (06/01/2012–09/17/2012 with a 14.8% increase in the S&P500 Index and 08/26/2010–02/18/2011 with a 28.3% increase in the S&P500 Index) and the two bottom diagrams correspond to bearish subperiods (07/01/2011–08/10/2011 with a 16.3% decrease in the S&P500 Index and 09/19/2008–11/20/2008 with a 40.1% decrease in the S&P500 Index).

In the comparison of the top and bottom diagrams of Figs. 6 and 7, we remark two important structural properties. First, the overall

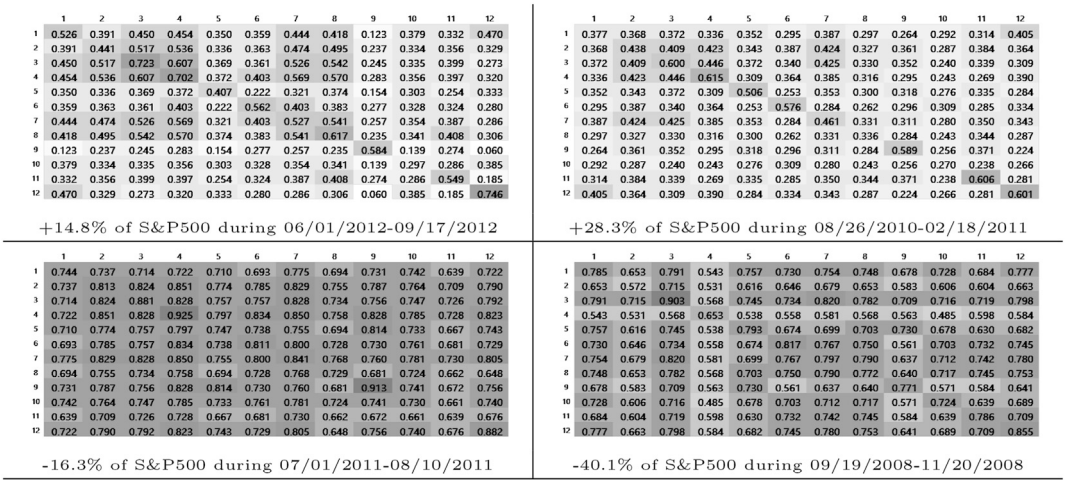

Fig. 6. Matrix representation of connectedness for the selected subperiods. The two lower matrices (bearish subperiods) are generally darker (higher correlations) than the two upper matrices. More importantly to our study, the two lower matrices show little contrast between the diagonal and off-diagonal elements, indicating a less clear cluster structure than the two upper matrices.

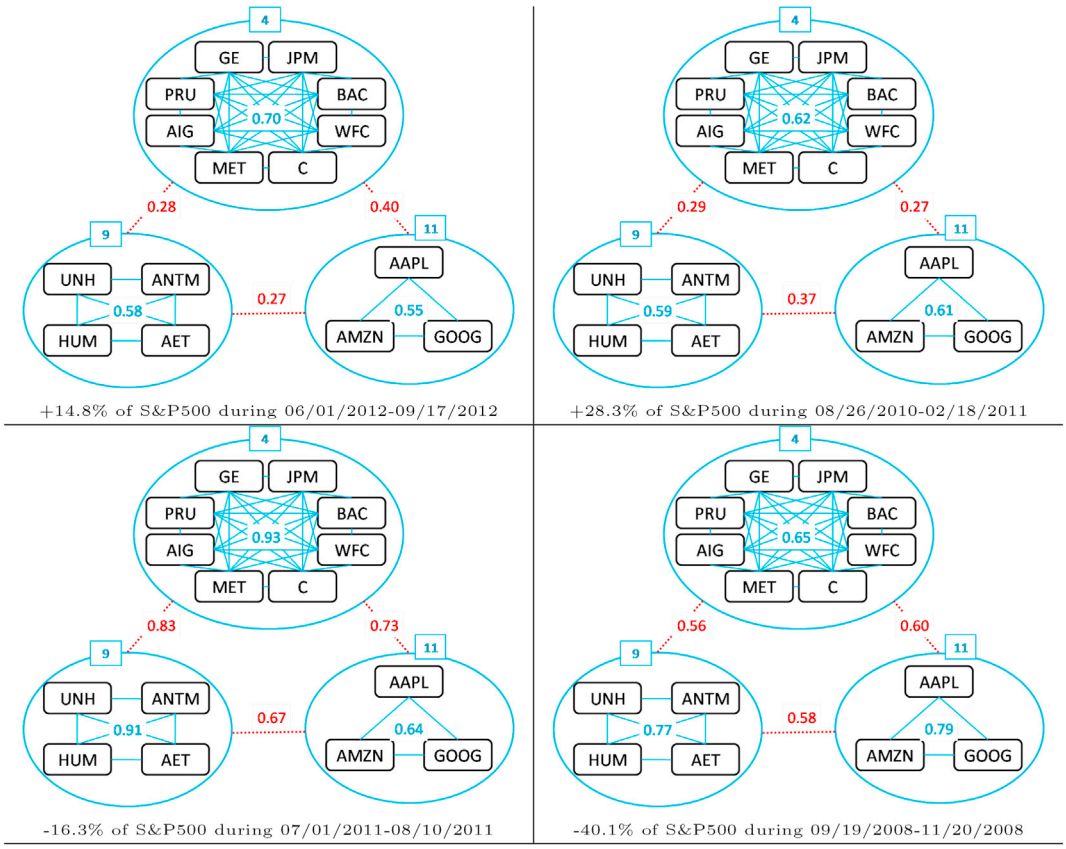

Fig. 7. A close-up view of the three subgroups from Fig. 6. The upper diagrams (bullish subperiods) have higher modularity than the lower diagrams (bearish subperiods).

level of correlation increases dramatically during the transition from a bullish period to a bearish period. In Fig. 6, the two lower matrices (bearish subperiods) are a lot darker (higher correlation) than the two upper matrices. In Fig. 7, the two lower diagrams also have much higher connectedness measures between and within subgroups. This phenomenon of being more correlated in a downturn has been pointed out by many practitioners and researchers, such as Buraschi et al. (2010) and Sandoval Jr. and Franca (2012). This

higher correlation in bear markets warns investors who use long-term historical correlations and believe that the correlations will remain at similar levels. The level of correlation changes over time, and the bad news is that the change is in the unfavorable direction at the time when investors desperately wish for moderate correlation to mitigate the effect of market-wide turmoil. If the quality of the investment opportunity set is assessed in terms of potential diversification benefits alone, this phenomenon implies the worsened quality of investable set in market downturn.

Second, and more importantly with regard to our study, the clustering tendency in a bearish subperiod is much weaker than that in a bullish subperiod. Although both INSC(P) and ITSC(P) simultaneously increase during the transition from a bullish to a bearish subperiod, the increase in ITSC(P) is much higher than that in INSC(P). As once moderate ITSC(P) increases rapidly, the relative difference between ITSC(P) and INSC(P) is narrowed. In other words, the once prominent clustered structure in normal or bullish markets is blurred in bearish markets. This phenomenon, which we call the *destruction of the cluster structure*, is noticeable from Fig. 6, where the contrast between the on- and off-diagonal elements in the top diagrams disappears in the bottom diagrams. This is also noticeable from Fig. 7, where the bottom diagrams display a less clear clustering tendency compared to the top diagrams. Stock investors allocate wealth into multiple industry sectors hoping that the diversification effects between the subgroups still hold in a downturn. However, another bad news in a downturn is that the widespread practice of sector diversification does not work very well. The worsened quality of the investment universe set with respect to sector diversification is assessed by our framework.

Should stressed markets lead to the destruction of the existing cluster structure, it may become necessary to identify a new cluster structure for understanding the investment opportunity set better. However, identifying a new cluster structure makes it difficult, or even impossible, to quantitatively compare the new cluster structure with the previous cluster structure. Our proposed connectedness measures allow us to quantify the level of deviations between the current cluster structure and the fixed cluster structure. The remarkable structural changes in the clustering tendency under different market conditions are captured through the proposed measures of INSC, ITSC, and modularity, of which modularity is the ultimate measure for quantifying the strength of the clustering tendency and of assessing the investment opportunity set in terms of the possible benefits of portfolio diversification.

4. Serving as an asset pricing factor: Is modularity factor priced?

Now that modularity is a meaningful indicator of the quality of the investment opportunity set and market fluctuations, the next question is whether past a co-movement tendency provides information that can generate a significant future return difference. This section investigates whether an individual security or a portfolio's sensitivity to the modularity can explain the expected return. The sensitivity, which we call "modularity beta'' hereafter, is defined as the coefficient of modularity in a multiple linear regression, where the other explanatory variables are the Fama-French factors[7] (Fama and French (1992, 1993; 2015)) (Section 4.1). We present estimation process for each stock's modularity beta (Section 4.2). We also present a two-way analysis of portfolio returns sorted by market beta and modularity beta (Section 4.3). We construct annually updated decile portfolios sorted by the modularity beta. (Section 4.4). The excess returns of the decile portfolios are then regressed on the established one-, three-, and five-factors. The varying non-zero intercepts across the decile portfolios substantiates the existence of the modularity factor.

4.1. Model

We define β_i^{MOD} as the coefficient of the time series of modularity factor, MOD_t, in the following multiple linear regression.

$$r_{i,t} - r_{f,t} = \beta_i^0 + \beta_i^{MOD} MOD_t + \beta_i^M MKT_t + \beta_i^S SMB_t + \beta_i^H HML_t + \beta_i^R RMW_t + \beta_i^C CMA_t + \varepsilon_{i,t}, \tag{6}$$

where $r_{i,t}$ denotes the return on stock i; $r_{f,t}$ is the risk-free return; MKT_t, SMB_t, HML_t, RMW_t, and CMA_t are the Fama-French five-factors (market, size, growth, profitability, and investment); and β_i^M, β_i^S, β_i^H, β_i^R, and β_i^C are the corresponding coefficients. Since the modularity time series MOD_t is not measured in portfolio returns, the intercept term β_i^{MOD} possesses a different meaning than alpha, which is an abnormal return in general linear factor asset pricing models.

4.2. Estimation of modularity betas

For each US stock, its modularity beta in (6) is estimated. Modularity sensitivity-sorted portfolios are accordingly constructed for 1992–2015. The construction for the modularity time series, the estimation for the modularity beta, and the construction of the modularity beta sorted portfolios use only data available as of the formation date.

Step 1. Set the year T = 1992.
Step 2. Collect historical return data for all stocks. Historical stock returns are obtained from the Center for Research in Security Prices (CRSP). Historical data for the Fama-French factors are collected from the Kenneth R. French Data Library. We use all common shares (CRSP share codes 10 and 11) actively traded on the New York Stock Exchange, American Stock Exchange, and NASDAQ.

[7] The momentum factor by Jegadeesh and Titman (1993) and Carhart (1997), quite powerful and widely accepted pricing factor, was also tested, and test results were overall similar. We present the results with the five-factors gathered from the data library of Prof. French. (*http : / / mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html*)

Step 3. Collect the residuals from recent history. Regress excess returns over the risk-free rate of all stocks by using the Fama-French three-factors[8] for the 60 months[9] from year T-5 to year T-1. In this step, stocks with missing records, or stocks with prices that fell below $5 once or more during any of the 60 months are excluded from the analysis.

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i^M MKT_t + \beta_i^S SMB_t + \beta_i^H HML_t + \eta_{i,t} \tag{7}$$

Then, by using the estimated betas from (7), collect the residuals of each stock.

$$e_{i,t} = r_{i,t} - r_{f,t} - \widehat{\beta}_i^M MKT_t - \widehat{\beta}_i^S SMB_t - \widehat{\beta}_i^H HML_t \tag{8}$$

The residual $e_{i,t}$ thus contains both the alpha and noise term in (7), both of which are left unexplained by the Fama-French three-factors.

Step 4. Construct historical monthly modularity time series. The representative set for modularity construction is chosen as the publicly traded top 60 companies[10] on the Fortune 500 list published in year T-1 which sorts US companies by their annual operating revenue in year T-2. Compute a sample correlation matrix of the historical daily returns of the 60 stocks from year T-5 to year T-1 and apply the MMC algorithm on the sample correlation matrix to generate a cluster solution. Using this cluster solution, generate historical modularity for each month of the five-year period. Specifically, calculate sample correlation matrices for each month of the five years and generate monthly modularity with respect to the cluster solution.[11]

Step 5. Regress the residuals by the modularity to estimate the historical sensitivity. By using the data for the same 60 months, estimate the historical sensitivity of each stock i, ψ_i^1, by using the simple linear regression of (9):

$$e_{i,t} = \psi_i^0 + \psi_i^1 MOD_t + \pi_{i,t}, \tag{9}$$

The quantity ψ_i^1 serves as an estimated modularity beta of stock i for the next 12 months.

Step 6. Continue to the next year. After the modularity betas for year T are estimated, go to Step 1 and increase T by 1 and repeat Steps 2–5. Repeat until modularity betas for all stocks for all years are estimated.

4.3. Two-way (4 × 4) portfolios sorted by market beta and modularity beta

Our discussion in the previous section suggests the possibility of informational content in the modularity measure regarding market fluctuations. As much as an individual security's sensitivity to the market factor, termed the market beta, is an effective asset pricing variable, can we expect similar usage from the modularity beta? If so, what is the interaction effect between the market factor and modularity factor? This subsection sorts individual securities by pre-rankings of market beta and modularity beta in order to address these questions.

Table 2 and Table 3 present the average monthly returns and post-market beta for two-way classified portfolios, value-weighted and equal-weighted, respectively. Panel A of both tables reveal a few interesting observations on the returns. First, stocks with low modularity betas tend to yield larger expected returns. Second, modularity beta sorting generates return differences of 2.89% (value-weighted) and 3.17% (equal-weighted) between both ends of classification. This dispersion is a little less compared to the return differences of 3.39% (value-weighted) and 5.82% (equal-weighted) that market beta sorting generates. Third, the effect of modularity beta sorting is consistent across the different levels of market betas. For all cases, where market beta is very high, high, low, or very low, the stocks with lower modularity betas generate higher returns. Panel B of both tables show that the post market betas of low modularity beta stocks and high modularity beta stocks are not very different, meaning that the source of return difference between low modularity beta stocks and high modularity beta stocks is not likely the market factor.

Why do modularity beta generate return differentials, and why do low modularity beta stocks yield higher expected returns? Note that the market factor is the weighted average of the stock return of the universe, and the modularity factor is the strength of the cluster structure. The two factors are related in that the strongly clustered structure means they are well segmented, hence a more stable market structure. Despite this general tendency, the results presented in Tables 2 and 3 suggest that modularity factor measures structural change that market factor is unable to capture. Low modularity beta stocks are less sensitive to structural changes and this robustness

[8] We use only three-factors for the residual estimations as in Paster and Stambaugh (2003). We also tested with the five-factors, and the results were similar.

[9] Indeed, we use 36 months of history for year T = 1992, 48 months for year T = 1993, and 60 months for all the other years thereafter.

[10] We use the Fortune list published in year T-1 because the list published in year T is generally unavailable at the beginning of year T. The number of companies is set to 60 to balance the efficiency of the algorithm and broadness of the set. Although including more than 60 stocks may enhance the broadness of the representative set, it may result in poor computational efficiency of the clustering algorithm and relative insufficiency of the number of samples over the number of parameters to be estimated for the correlation matrix.

[11] For each month, there are only 18–23 trading days. This number of observations is statistically insufficient to guarantee the stability of the statistical estimators generated on the correlation matrix. To enhance the stability of the modularity time-series, we use the three-month moving average smoothing of original monthly observed modularity. Our tests with different lengths of smoothing, two-, four-, five-, and six-months, show similar results as well.

Table 2
Two-way value-weighted portfolios returns sorted by market beta and modularity beta (January 1992–December 2015).

Panel A: Average Returns (%, per annum)					
	All	Low-β^{MOD}	$\beta^{MOD}-2$	$\beta^{MOD}-3$	High-β^{MOD}
All	11.07	12.89	11.17	10.22	10.00
Low-β^{MKT}	9.03	10.76	8.70	8.37	7.93
$\beta^{MKT}-2$	11.08	12.06	10.93	10.32	10.70
$\beta^{MKT}-3$	11.74	13.09	12.21	11.33	10.58
High-β^{MKT}	12.42	14.31	13.72	11.76	11.59
Panel B: Post β^{MKT}					
	All	Low-β^{MOD}	$\beta^{MOD}-2$	$\beta^{MOD}-3$	High-β^{MOD}
All	1.05	1.08	0.96	0.99	1.18
Low-β^{MKT}	0.69	0.72	0.63	0.66	0.78
$\beta^{MKT}-2$	0.88	0.86	0.85	0.87	0.95
$\beta^{MKT}-3$	1.12	1.14	1.09	1.11	1.15
High-β^{MKT}	1.53	1.54	1.52	1.48	1.56

Table 3
Two-way equal-weighted portfolios returns sorted by market beta and modularity beta (January 1992–December 2015).

Panel A: Average Returns (%, per annum)					
	All	Low-β^{MOD}	$\beta^{MOD}-2$	$\beta^{MOD}-3$	High-β^{MOD}
All	8.10	9.79	8.12	7.71	7.44
Low-β^{MKT}	5.70	6.52	6.73	4.87	3.35
$\beta^{MKT}-2$	7.95	9.39	6.53	7.80	8.92
$\beta^{MKT}-3$	9.06	9.73	9.55	9.42	7.98
High-β^{MKT}	9.49	12.34	10.43	10.13	7.83
Panel B: Post β^{MKT}					
	All	Low-β^{MOD}	$\beta^{MOD}-2$	$\beta^{MOD}-3$	High-β^{MOD}
All	1.11	1.16	1.06	1.06	1.23
Low-β^{MKT}	0.68	0.72	0.60	0.67	0.85
$\beta^{MKT}-2$	0.88	0.83	0.91	0.85	0.97
$\beta^{MKT}-3$	1.13	1.17	1.12	1.13	1.13
High-β^{MKT}	1.58	1.59	1.62	1.51	1.58

leads to the higher expected returns. Although we find little possible explanation that stocks with low modularity should be considered riskier, these stocks yield higher returns. Our results indicate that the modularity factor is more likely to be an anomaly factor than a risk-return trade-off factor.

If the modularity beta generates return differentials, then a naturally following question is which stocks have high/low modularity betas. We find that the modularity beta is not a constant property of a stock. In the two way classification of stocks presented in Tables 2 and 3, we find that stocks once in the lowest modularity beta basket may move to the highest modularity beta basket in the next year, or vice versa. The time variation of an individual stock's modularity may explain why the return differentials created by modularity beta sorting is not captured by existing famous asset pricing factors (to be presented in the next subsection). That is, if modularity beta were a constant property of a stock, then the property could have been identified by traditional approaches based on the characteristics of individual firms, such as stock prices or financial ratios (Fama and French (1992, 1993; 2015); Carhart (1997); Jegadeesh and Titman (1993)).

4.4. Decile portfolio tests: are there significant return differences?

Our hypothesis is that decile portfolios sorted by sensitivity to modularity series should display a systematic return difference that is not explained by other asset pricing models. Based on the ranking of $\widehat{\psi}_i^1$ from (9), we construct value-weighted and equal-weighted decile portfolios. $\widehat{\psi}_i^1$ serves as the "predicted modularity beta'' of stock i for the next 12 months.[12] The top (bottom) decile portfolio contains the stocks with the lowest (highest) predicted beta. In case of a missing record in a stock return, we assume that the stock returns for the missing month and all the following months are all equal to zero.

Table 4 and Table 5 present test results regarding the returns of decile portfolios, value-weighted and equal-weighted, respectively. In Panel A, the first row presents the annualized returns and standard deviations of the decile portfolios. The last column corresponds to

[12] Paster and Stambaugh (2003) not only estimated the historical sensitivity but also forecasted the sensitivity with the other characteristics of each firm, and the result with forecast sensitivity was slightly better. In this study, we simply set the historical beta as the predicted beta because we have no plausible candidate or conjecture with regard to which firm characteristics predict future modularity beta well.

Table 4
Decile portfolios formed by modularity-beta (value-weighted) (January 1992–December 2015).

Panel A: Returns and alphas with respect to the 1-,3-,5-factor models (%, per annum)											
	Low-β^{MOD}	2	3	4	5	6	7	8	9	High-β^{MOD}	1-10 (Low-High)
Return	13.37	7.75	7.99	9.68	8.65	7.37	8.58	7.42	8.36	6.95	6.42
Std. Dev.	20.22	15.08	13.98	13.17	14.20	14.52	15.77	16.06	17.07	20.64	15.44
CAPM alpha	2.71	−1.35	−0.65	1.28	−0.26	−1.69	−1.13	−2.35	−1.89	−4.66	7.38
(t-statistics)	(1.11)	(-0.9)	(-0.47)	(1.04)	(-0.2)	(-1.29)	(-0.85)	(-1.66)	(-1.29)	(-2.34)	(2.33)
3-Factor alpha	3.26	−1.24	−1.60	0.59	−1.09	−1.99	−1.40	−3.09	−2.52	−4.96	8.21
(t-statistics)	(1.35)	(-0.82)	(-1.31)	(0.56)	(-0.9)	(-1.55)	(-1.13)	(-2.49)	(-1.78)	(-2.48)	(2.6)
5-Factor alpha	4.90	−2.24	−3.13	−0.99	−2.12	−2.57	−1.88	−3.66	−3.53	−5.56	10.47
(t-statistics)	(1.96)	(-1.43)	(-2.53)	(-0.94)	(-1.71)	(-1.92)	(-1.45)	(-2.82)	(-2.4)	(-2.7)	(3.2)
Panel B: Exposure to the five-factors											
	Low-β^{MOD}	2	3	4	5	6	7	8	9	High-β^{MOD}	1-10 (Low-High)
β_{MKT}	0.99	0.95	0.95	0.92	0.94	0.94	1.04	1.07	1.13	1.25	−0.26
(t-statistics)	(18.06)	(27.65)	(34.99)	(39.29)	(34.78)	(31.95)	(36.64)	(37.73)	(34.94)	(27.57)	(-3.62)
β^{SMB}	0.15	−0.07	−0.10	−0.14	−0.03	−0.04	−0.15	−0.17	−0.04	0.17	−0.03
(t-statistics)	(2.05)	(-1.57)	(-2.82)	(-4.53)	(-0.81)	(-1.14)	(-4.15)	(-4.51)	(-1.03)	(2.94)	(-0.29)
β^{HML}	−0.01	−0.08	0.12	0.05	0.08	0.07	0.09	0.18	0.08	0.05	−0.06
(t-statistics)	(-0.06)	(-1.28)	(2.56)	(1.23)	(1.75)	(1.45)	(1.85)	(3.79)	(1.45)	(0.64)	(-0.45)
β^{RMW}	−0.16	0.14	0.19	0.17	0.08	0.13	0.09	0.09	0.14	0.20	−0.36
(t-statistics)	(-1.57)	(2.09)	(3.73)	(3.82)	(1.63)	(2.36)	(1.71)	(1.75)	(2.25)	(2.34)	(-2.68)
β^{CMA}	−0.29	0.11	0.20	0.26	0.22	−0.03	0.00	0.03	0.11	−0.15	−0.14
(t-statistics)	(-2.13)	(1.28)	(2.92)	(4.45)	(3.2)	(-0.46)	(0.03)	(0.41)	(1.37)	(-1.34)	(-0.78)
Panel C: Return decomposition by the five-factors											
	Low-β^{MOD}	2	3	4	5	6	7	8	9	High-β^{MOD}	1-10 (Low-High)
Excess return (p.a.)	10.77	5.15	5.38	7.07	6.04	4.76	5.97	4.81	5.76	4.34	6.42
α	4.90	−2.24	−3.13	−0.99	−2.12	−2.57	−1.88	−3.66	−3.53	−5.56	10.47
$\beta^{MKT}*\overline{MKT}$	7.20	6.89	6.92	6.67	6.87	6.82	7.57	7.81	8.21	9.09	−1.89
$\beta^{SMB}*\overline{SMB}$	0.36	−0.17	−0.25	−0.34	−0.07	−0.11	−0.38	−0.41	−0.11	0.43	−0.07
$\beta^{HML}*\overline{HML}$	−0.02	−0.25	0.39	0.16	0.27	0.24	0.29	0.60	0.26	0.16	−0.18
$\beta^{RMW}*\overline{RMW}$	−0.63	0.52	0.74	0.65	0.32	0.51	0.36	0.36	0.53	0.77	−1.40
$\beta^{CMA}*\overline{CMA}$	−1.04	0.39	0.71	0.93	0.77	−0.12	0.01	0.10	0.40	−0.54	−0.50

Table 5
Decile portfolios formed by modularity-beta (equally-weighted) (January 1992–December 2015).

Panel A: Returns and alphas with respect to the 1-,3-,5-factor models (%, per annum)											
	Low-β^{MOD}	2	3	4	5	6	7	8	9	High-β^{MOD}	1-10 (Low-High)
Return	13.46	12.49	11.35	11.24	11.00	10.41	9.78	10.61	11.08	9.20	4.26
Std. Dev.	17.36	14.98	13.98	13.37	13.45	13.36	13.71	14.68	15.67	19.33	9.92
CAPM alpha	3.72	3.57	2.85	2.96	2.70	2.01	1.16	1.54	1.58	−1.82	5.53
(t-statistics)	(1.88)	(2.22)	(1.9)	(2.11)	(1.89)	(1.52)	(0.89)	(1.12)	(1.08)	(-0.96)	(2.8)
3-Factor alpha	1.61	1.11	0.35	0.63	0.28	−0.26	−1.05	−0.73	−0.62	−3.74	5.34
(t-statistics)	(1.24)	(1.1)	(0.38)	(0.72)	(0.33)	(-0.33)	(-1.34)	(-0.82)	(-0.66)	(-2.63)	(2.69)
5-Factor alpha	1.19	−0.23	−1.12	−0.77	−1.16	−1.58	−2.12	−1.74	−1.56	−3.93	5.12
(t-statistics)	(0.9)	(-0.23)	(-1.25)	(-0.89)	(-1.39)	(-2.07)	(-2.73)	(-1.97)	(-1.63)	(-2.64)	(2.48)
Panel B: Exposure to the five-factors											
	Low-β^{MOD}	2	3	4	5	6	7	8	9	High-β^{MOD}	1-10 (Low-High)
β_{MKT}	0.94	0.91	0.88	0.85	0.85	0.86	0.88	0.94	0.98	1.12	−0.19
(t-statistics)	(32.05)	(42.66)	(45)	(44.98)	(46.72)	(51.44)	(51.84)	(48.71)	(46.63)	(34.43)	(-4.12)
β^{SMB}	0.71	0.55	0.45	0.41	0.43	0.38	0.37	0.38	0.46	0.54	0.17
(t-statistics)	(18.63)	(19.58)	(17.78)	(16.85)	(18.08)	(17.49)	(16.61)	(14.9)	(16.77)	(12.77)	(2.83)
β^{HML}	0.28	0.34	0.36	0.33	0.32	0.32	0.33	0.35	0.30	0.25	0.03
(t-statistics)	(5.51)	(9.37)	(10.69)	(10.06)	(10.31)	(11.09)	(11.39)	(10.63)	(8.35)	(4.41)	(0.38)
β^{RMW}	0.13	0.26	0.26	0.24	0.22	0.22	0.19	0.19	0.15	0.02	0.11
(t-statistics)	(2.29)	(6.45)	(7.1)	(6.65)	(6.28)	(6.79)	(5.81)	(5.16)	(3.89)	(0.3)	(1.26)
β^{CMA}	−0.09	−0.00	0.04	0.06	0.12	0.08	0.04	0.02	0.05	0.03	−0.12
(t-statistics)	(-1.17)	(-0.04)	(0.89)	(1.38)	(2.57)	(1.83)	(0.91)	(0.31)	(1.03)	(0.42)	(-1.06)
Panel C: Return decomposition by the five-factors											
	Low-β^{MOD}	2	3	4	5	6	7	8	9	High-β^{MOD}	1-10 (Low-High)
α	1.19	−0.23	−1.12	−0.77	−1.16	−1.58	−2.12	−1.74	−1.56	−3.93	5.12
$\beta^{MKT}*\overline{MKT}$	6.82	6.65	6.41	6.17	6.19	6.29	6.44	6.88	7.14	8.18	−1.36
$\beta^{SMB}*\overline{SMB}$	1.75	1.34	1.11	1.02	1.05	0.94	0.91	0.93	1.13	1.33	0.41
$\beta^{HML}*\overline{HML}$	0.90	1.12	1.17	1.06	1.05	1.04	1.09	1.15	0.98	0.81	0.10
$\beta^{RMW}*\overline{RMW}$	0.49	1.01	1.02	0.92	0.84	0.83	0.73	0.73	0.60	0.07	0.42
$\beta^{CMA}*\overline{CMA}$	−0.31	−0.01	0.15	0.23	0.42	0.27	0.14	0.05	0.19	0.12	−0.43

the difference between the decile 1 portfolio (stocks with low modularity betas) and the decile 10 portfolio (stocks with high modularity betas). This "1–10" spread portfolio is equivalent to a net zero investment portfolio in which an investor buys the first decile and sells the last decile by the same amount of wealth. In the second to the fourth row, the levels of alphas with respect to the factor models are presented. Specifically, R^j is the excess return of decile portfolio $j = 1, 2, .., 10$, and $1 - 10$. Then, α_j from (10) are presented:

$$R_t^j = \alpha_j + B_j F_t + \eta_{j,t} \tag{10}$$

In (10), F_t is a vector of the one-, three-, or five-factors in month t. B_j is the sensitivity vector with respect to the factors. As shown in the last column of Panel A, the "1–10" portfolio has an annualized return of 6.42% (equal weighted case has 4.26%), and the α_{1-10} is significantly positive with respect to all of the three types of Fama-French factor model variants. Hence, the annualized alphas with respect to the variants are in the range of 7.38%–10.42% (the equally weighted case is in the range of 5.12%–5.53%).

We also test the hypothesis that all alphas are jointly equal to zero by adopting the method of Gibbons et al. (1989). For the null hypothesis of $\alpha_1 = \alpha_2 = \cdots = \alpha_{10} = 0$, we find that this hypothesis is rejected at the 1% significance level for all CAPM, three-factor, and five-factor alphas in the cases of both value-weighted and equal-weighted portfolios. Panel B presents B_j in (10) to confirm that the differences in returns are not captured by the Fama-French factors. The last column in Panel B of Table 4 indicates that the "1–10" portfolio is statistically tilted toward the firms that have low market betas and that are less profitable. Thus, the existing factor model further strengthens the evidence of the return differences from the modularity sensitivity-sorted portfolios. The last column in Panel B of Table 5 indicates the "1–10" portfolio is statistically tilted toward the firms that have low market beta and that are small. However, this tilt does not breach the statistical significance of return difference.

As an extension to Panel B, Panel C decomposes the returns of the decile portfolios with respect to the exposures for each of the five-factors. The decomposition is performed by multiplying the estimated betas in Panel B by the historical average of the corresponding factors. Specifically, by taking the average of both sides in (10) over all months, we have

$$\overline{R^j} = \alpha_j + \beta_j^{MKT}\overline{MKT} + \beta_j^{SMB}\overline{SMB} + \beta_j^{HML}\overline{HML} + \beta_j^{RMW}\overline{RMW} + \beta_j^{CMA}\overline{CMA}, \tag{11}$$

where $\overline{factor}$ is the historical average of the factor portfolio return. Panel C indicates that some significant exposure of the "1–10" portfolio toward the five-factors even strengthens the return difference of 6.42% to the five-factor alpha of 10.47% (or, 4.26%–5.12% in the equally weighted case).

5. Enhancing the investment opportunity set with the modularity factor

If asset pricing factors effectively summarize the investment opportunity set, factor investors only need to allocate their wealth into factor-mimicking basis portfolios. This section includes a modularity factor portfolio to the set of the Fama-French factor basis portfolios and investigates whether the inclusion potentially benefits factor investors by expanding their investment opportunity sets. Since modularity factor itself is not tradeable portfolio yet, each of following subsections adopts different mimicking strategies to generate a tradeable modularity factor basis portfolio.

4.5. *Modularity basis portfolios: The difference between the extreme decile portfolios*

Unlike the other Fama-French factors, the time series of modularity itself is not a tradeable portfolio. The difference between the extreme decile portfolios presented above is the easily implementable factor portfolio that mimics the modularity factor. We notate this annually updating modularity factor portfolio as MOD_s, where the subscript S implies the spread between the two extreme portfolios. The following ex-post analysis for the 24-year period from January 1992 to December 2015 uses the standard mean-variance tangent portfolio construction,[13] similar to that employed by Paster and Stambaugh (2003). Specifically, the historical monthly returns of MOD_s and the five-factors are considered to construct tangent portfolios. The value-weighted and equal-weighted cases are both considered, notated as $\text{MOD}_\text{s}(\text{VW})$ and $\text{MOD}_\text{s}(\text{EW})$, respectively.

Table 6 presents the ex-post mean-variance efficient portfolios with several combinations of the five-factors and MOD_s. Investing only in the MKT factor, equivalent to the one-factor model (i.e., CAPM), would result in a monthly Sharpe ratio of 0.142. In the case that an investor adds another factor to the MKT factor, adding $\text{MOD}_\text{s}(\text{VW})$ or $\text{MOD}_\text{s}(\text{EW})$ would result in a monthly Sharpe ratio of 0.167 or 0.206, generally no worse than adding SMB or HML. Overall, Table 6 suggests that both $\text{MOD}_\text{s}(\text{VW})$ and $\text{MOD}_\text{s}(\text{EW})$ are attractive basis portfolios to be included.

Table 7 presents the cross-correlation matrix for the basis factors. Both $\text{MOD}_\text{s}(\text{VW})$ and $\text{MOD}_\text{s}(\text{EW})$ are highly independent of the other factors, as suggested in Tables 4 and 5. Despite being a net-zero investment portfolio, MOD_s has a positive annualized return and is relatively uncorrelated with MKT. This makes MOD_s an attractive candidate to be added to portfolios that are highly correlated to the MKT factor, such as market index funds.

Table 8 presents the enhanced performance of a market index portfolio when MOD_s are added. Adding 10% of the $\text{MOD}_\text{s}(\text{VW})$ exposure to the MKT factor would increase the annualized return while sustaining a similar standard deviation. The monthly Sharpe

[13] The ex-post optimal portfolio is obtained from $w = \frac{\Sigma^{-1}(r - r_f)}{1^t\Sigma^{-1}(r - r_f)}$, where r is the historical average return vector, Σ is the historical covariance matrix of the factor portfolios under consideration, and r_f is the risk-free rate.

Table 6
Weights in the ex-post tangency portfolio and the monthly Sharpe ratios (January 1992–December 2015). Effects of adding modularity replicating portfolio MOD_s to basis assets can expand investment opportunity set, resulting in better ex-post Sharpe ratios.

Number of	MKT	SMB	HML	RMW	CMA	MOD_s	MOD_s	Sharpe Ratio
Instruments						(Value-Weighted)	(Equally-Weighted)	(monthly)
1	100							0.142
2	73.55	26.45						0.147
2	50.49		49.51					0.187
2	64.62					35.38		0.167
2	42.87						57.13	0.206
3	38.78	18.13	43.1					0.195
4	24.46	22.56	5.34	47.64				0.295
4	30.62	5.38	−2.17		66.17			0.251
4	31.93	13.37	36.96			17.74		0.221
4	29.88	9.72	26.23				34.17	0.240
5	22.11	13.83	−15.55	38.63	40.97			0.362
6	20.23	12.15	−13.48	35.95	36.86	8.27		0.397
6	20.87	10.43	−15.09	32.11	37.94		13.74	0.395

Table 7
Cross-correlation matrix of the five-factors including MOD_s (January 1992–December 2015). Correlation of modularity replicating portfolios are negligibly correlated to each of the five-factors.

	MKT	SML	HML	RMW	CMA	MOD_s(VW)	MOD_s(EW)
MKT	1	0.21	−0.23	−0.44	−0.35	−0.05	−0.21
SML		1	−0.19	−0.50	−0.04	0.04	0.01
HML			1	0.45	0.67	−0.11	0.09
RMW				1	0.21	−0.15	0.13
CMA					1	−0.10	0.01
MOD_s(VW)						1	0.64
MOD_s(EW)							1

Table 8
Performance of the enhanced market index portfolio (Annualized, January 1992–December 2015). Because modularity replicating portfolios have positive returns and low market exposure, mixing it to market portfolio demonstrates enhancing performance.

	Return (%)	Std. Dev. (%)	Sharpe Ratio
MKT	9.89	14.78	0.493
MKT+10% MOD_s(VW)	10.35	14.79	0.524
MKT+10% MOD_s(EW)	10.28	14.60	0.525
MKT+20% MOD_s(VW)	10.81	14.98	0.548
MKT+20% MOD_s(EW)	10.66	14.49	0.556

ratio is enhanced from 0.493 to 0.524. Adding MOD_s(EW) also yields similar results. Fig. 8 displays the scenarios of cumulative wealth growth for the investment horizon.

4.6. Modularity basis portfolio: The minimum idiosyncratic risk procedure by Lehmann and Modest (2005)

While the spread portfolio, MOD_s, is a legitimate modularity factor-mimicking portfolio, considering only the extreme decile portfolios may not fully reflect the effect of the newly tested factor. While the Fama and MacBeth (1973) mimicking portfolio is a classic alternative, Lehmann and Modest (2005) pointed out that there is no guarantee that the Fama-Macbeth factor-mimicking portfolio will have sufficiently high correlation to the original factor within finite samples. This subsection adopts the minimum idiosyncratic risk procedure proposed by Lehmann and Modest (2005) to construct another factor-mimicking portfolio. This procedure mimics a factor by taking the difference between the market-wide equally weighted portfolio and the portfolio orthogonal to the factor.[14]

We notate the factor-mimicking portfolio as MOD_p, where the subscript P implies portfolio. To present sensible numbers, MOD_p is

[14] The orthogonal portfolio w_{orth} to the new factor solves the following optimization problem: $(P)\ min\ w'_{orth}w_{orth}\ s.t.\ w'_{orth}1 = 1;\ w'_{orth}\boldsymbol{\beta} = 0$, where $\mathbf{1}$ is a vector of ones and $\boldsymbol{\beta}$ is obtained by the factor linear regression. The solution to this problem is given as $w_{orth} = \frac{1}{N}\left[\frac{\sigma_\beta^2+\bar{\beta}^2}{\sigma_\beta^2}1 - \frac{\bar{\beta}}{\sigma_\beta^2}\boldsymbol{\beta}\right]$, where N is the number of stocks used to construct the portfolio, σ_β is the standard deviation of $\boldsymbol{\beta}$, and $\bar{\beta}$ is the mean of $\boldsymbol{\beta}$. The factor mimicking portfolio is therefore $w_{mimic} = \frac{1}{N}1 - w_{orth} = \frac{\bar{\beta}}{N\sigma_\beta^2}(\boldsymbol{\beta} - 1\bar{\beta})$.

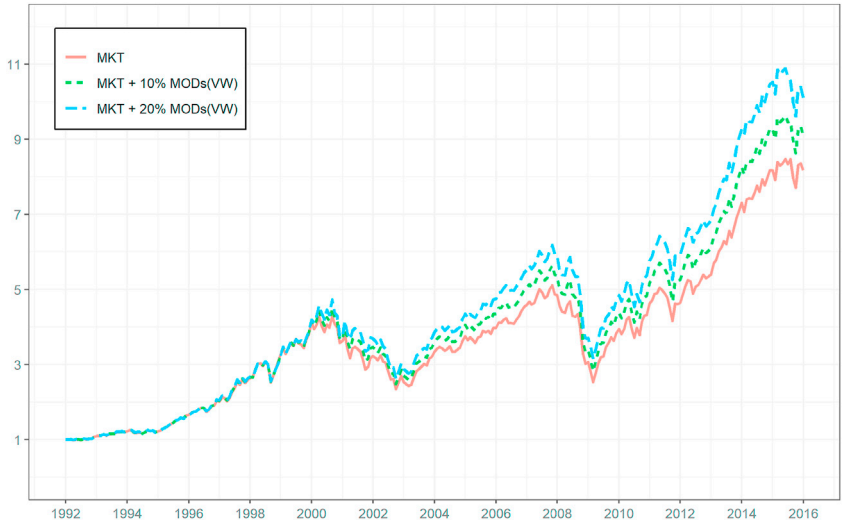

Fig. 8. Cumulative wealth growth of the enhancing scenarios by.MOD_S

Table 9
Weights in the ex-post tangency portfolio and monthly Sharpe ratios (January 1992–December 2015). Effects of adding modularity replicating portfolio MOD_p to basis assets can expand investment opportunity set, resulting in better ex-post Sharpe ratios.

Number of	MKT	SMB	HML	RMW	CMA	MOD_p	Sharpe Ratio
Instruments							(monthly)
1	100						0.142
2	73.55	26.45					0.147
2	50.49		49.51				0.187
2	57.36					42.64	0.199
3	38.78	18.13	43.1				0.195
4	24.46	22.56	5.34	47.64			0.295
4	30.62	5.38	−2.17		66.17		0.251
4	33.31	16.55	28.17			21.97	0.234
5	22.11	13.83	−15.55	38.63	40.97		0.362
6	21.84	12.15	−17.34	31.84	43.39	8.12	0.387

scaled to have the same ex-post standard deviation as that of MOD_s(VW). Table 9–Table 11 and Fig. 9 repeat the same analysis as that discussed in Section 5.1. The correlation between MOD_s and MOD_p is 0.725, and the overall results are similar. MOD_p is also an attractive instrument with regard to the mean-variance method (Table 9). The low correlation of MOD_p with respect to the market portfolio facilitates its usage as a portfolio overlaying the market index portfolios (Table 10, Table 11, and Fig. 9). We conclude that both

Table 10
Cross-correlation matrix of the five-factors, including MOD_p (January 1992–December 2015). Correlation of modularity replicating portfolios are negligibly correlated to each of the five-factors.

	MKT	SML	HML	RMW	CMA	MOD_p
MKT	1.00	0.21	−0.23	−0.44	−0.35	−0.23
SML		1.00	−0.19	−0.50	−0.04	−0.15
HML			1.00	0.45	0.67	0.18
RMW				1.00	0.21	0.35
CMA					1.00	−0.04
MOD_p						1.00

Table 11
Performance of the enhanced market index portfolio (Annualized, January 1992–December 2015).

	Return (%)	Std. Dev. (%)	Sharpe Ratio
MKT	9.89	14.78	0.493
MKT+10% MOD_p	10.47	14.49	0.543
MKT+20% MOD_p	11.05	14.39	0.587

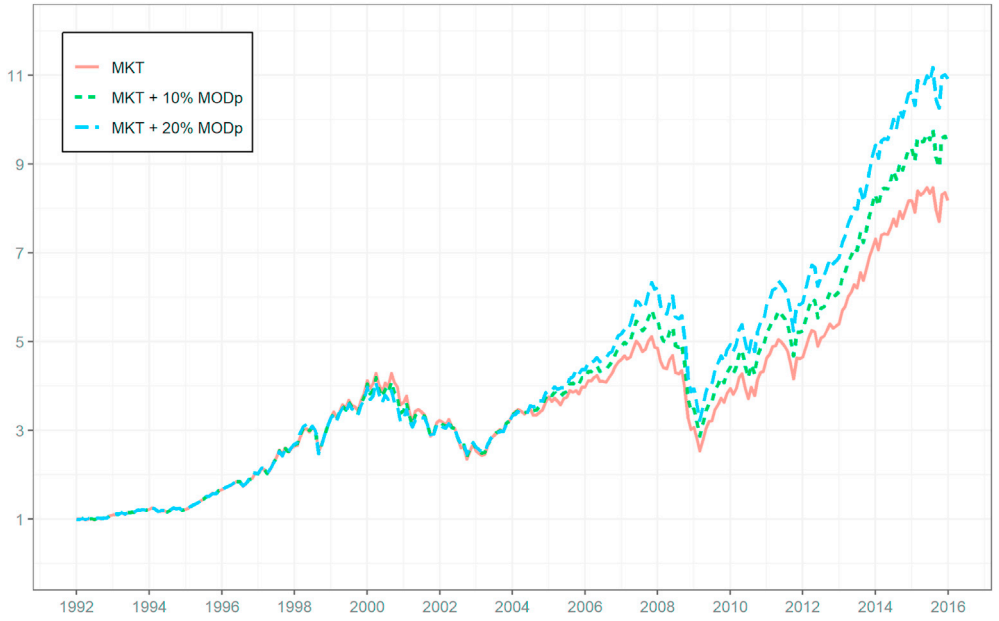

Fig 9. Cumulative return on the enhancement scenarios with.MOD_p

mimicking methods demonstrate similarly enhanced performance for factor-based investments. Thus, the inclusion of the modularity factor expands the investment opportunity set.

5. Concluding remarks

We propose a modularity measure that quantifies the strength of a cluster structure in financial assets. Upon a given cluster structure, the modularity is the difference between the INSC (inner-sector connectedness) and the ITSC (inter-sector connectedness). This measure is built on a correlation matrix, but taking the difference between the two quantities allows it to be somewhat independent with respect to the original correlation matrix (e.g., average market-wide correlations by Pollet and Wilson (2010)).

On a market-wide level, the modularity contrasts in a few bullish and bearish subperiods, demonstrating that a bearish market is characterized by low modularity, thereby losing the potential benefit of portfolio diversification. The empirical results using modularity-beta sorted portfolios demonstrate that the modularity measure is indeed a valid risk factor driving asset returns. Stocks with low sensitivity to the modularity factor have considerably higher expected returns, even after accounting for exposure to the Fama-French three- or five-factors. Further, the difference between extreme decile portfolios or the mimicking method of Lehmann and Modest (2005) creates modularity factor portfolios that can be used to enlarge the investment opportunity set for passive investors.

Grouping stocks into clusters has been a popular subject in empirical studies, and studies find that the grouping behavior of stocks varies by country. For example, the US market is known to be clustered by industry, but the Japanese market is less clustered by industry. A few developing countries have stocks clustered by conglomerative capitals as well. The presented framework is immunized to such country specific characteristics, because it approaches a clustered structure with the co-movement behavior of stock returns. Future studies could apply this framework to other countries than the US.

A future extension of this study could include agglomerating the current construction of modularity to statistical techniques such as principal component analysis or hidden graphical models. The latent structure of the financial market may be further refined into an alternative form of the modularity factor. Although this study uses daily stock market data in the analysis, analyzing the time-varying cluster property using our framework could be applied to many other types of financial data, including high-frequency trading data, fixed income markets data, and foreign exchange market data.

Conflict of interest

We have no conflict of interest to declare. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.iref.2020.09.004.

References

Ahn, D., Conrad, J., & Dittmar, R. (2009). Basis assets. *Review of Financial Studies, 22*, 5133–5174.

Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics, 104*(3), 535–559.

Buraschi, A., Porchia, P., & Trojan, F. (2010). Correlation risk and optimal portfolio choice. *The Journal of Finance, LXV*, 393–420.

Carhart, M. (1997). On persistence in mutual fund performance. *The Journal of Finance, 52*, 57–82.

Chandrasekaran, V., Parrilo, P., & Willsky, A. (2012). Latent variable graphical model selection via convex optimization. *Annals of Statistics, 40*, 1935–1967.

Diebold, F. X., & Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics, 182*, 119–134.

Fama, E., & French, K. (1992). The cross-section of expected stock returns. *The Journal of Finance, 47*, 427–465.

Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*, 3–56.

Fama, E., & French, K. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*, 1–22.

Fama, E., & MacBeth, J. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy, 81*, 607–636.

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika, 10*, 507–521.

Fisher, R. (1924). The distribution of the partial correlation coefficient. *Metron, 3*, 329–332.

Gibbons, M., Ross, S., & Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica, 57*, 1121–1152.

Jang, W., Lee, J., & Chang, W. (2011). Currency crises and the evolution of foreign exchange market: Evidence from minimum spanning tree. *Physica A, 390*, 707–718.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance, 48*, 65–91.

Lehmann, B., & Modest, D. (2005). Diversification and the optimal construction of basis portfolios. *Management Science, 51*, 581–598.

Materassi, D., & Innocenti, G. (2009). Unveiling the connectivity structure of financial networks via high-frequency analysis. *Physica A, 388*, 3866–3878.

Newman, M. (2006). Modularity and community structure in networks. *PNA, 103*, 8573–8574.

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review, 69*.

Paster, L., & Stambaugh, R. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy, 111*, 642–685.

Pollet, J., & Wilson, M. (2010). Average correlation and stock market returns. *Journal of Financial Economics, 96*, 364–380.

Sandoval, L., Jr., & Franca, I. (2012). Correlation of financial market in times of crisis. *Physica A, 391*, 187–208.

Stone, E., & Ayroles, J. (2009). Modulated modularity clustering as an exploratory tool for functional genomic inference. *PLoS Genetics, 5*.

Zimmerman, D., Zumbo, B., & Williams, R. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica, 24*, 133–158.