Guo, Li; Tao, Yubo; Härdle, Wolfgang Karl

Working Paper

Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective

IRTG 1792 Discussion Paper, No. 2018-032

Provided in Cooperation with:
Humboldt University Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series"

Suggested Citation: Guo, Li; Tao, Yubo; Härdle, Wolfgang Karl (2018) : Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective, IRTG 1792 Discussion Paper, No. 2018-032, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/230743

IRTG 1792 Discussion Paper 2018-032

Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective

Li Guo *
Yubo Tao *
Wolfgang Karl Härdle *²

* Singapore Management University, Singapore
*² Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective

Li Guo*
Lee Kong Chian School of Business, Singapore Management University

Yubo Tao
School of Economics, Singapore Management University

Wolfgang Karl Härdle
Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin
Sim Kee Boon Institute for Financial Economics, Singapore Management University

Thursday 24th May, 2018

Abstract

In this paper, we study the latent group structure in cryptocurrencies market by forming a dynamic return inferred network with coin attributions. We develop a dynamic covariate-assisted spectral clustering method to detect the communities in dynamic network framework and prove its uniform consistency along the horizons. Applying our new method, we show the return inferred network structure and coin attributions, including algorithm and proof types, jointly determine the market segmentation. Based on the network model, we propose a novel "hard-to-value" measure using the centrality scores. Further analysis reveals that the group with a lower centrality score exhibits stronger short-term return reversals. Cross-sectional return predictability further confirms the economic meanings of our grouping results and reveal important portfolio management implications.

Keywords: Community Detection, Dynamic Network, Return Predictability, Behavioural Bias, Market Segmentation, Bitcoin.

*The authors gratefully acknowledge all the participants who attended the workshop "Crypto-Currencies in a Digital Economy" in Humboldt-Universität zu Berlin for helpful discussions and comments. Send correspondence to Li Guo at liguo.2014@pbs.smu.edu.sg.

1 Introduction

The invention of Bitcoin by Satochi Nakamoto (Nakamoto, 2008) in 2008 spurred the creation of many new cryptocurrencies known as *Altcoins*. As of April 18th, 2018, more than 800 cryptocurrencies are trading actively worldwide with more than $100 billion market capitalizations. The growing number of altcoins stimulates the investors to investigate the internal relationships between those altcoins and to make a fortune with it. Nevertheless, unlike equities market uses industry classification (GIC and SIC), we have no stringent criteria to classify the cryptocurrencies. Although many of them use similar cryptographic technologies, subtle differences in algorithmic designs or other characteristics may lead to complete different price trajectories. Due to the same reason, the fundamental characteristics of cryptocurrencies are hard to price, and thus makes it worthwhile to figure out how fundamental characteristics (e.g. algorithm and proof type) differentiate the performance of different cryptocurrencies.

As a natural question, one may wonder whether the same classification methodology used in equities market can be applied to cryptocurrencies. However, market segmentation of cryptocurrencies is a more complicated issue than that of equities in many aspects, and one of most severe issue is the data scarcity. For example, Hoberg and Phillips (2016) provide a new measure of product differentiation based on textual analysis of 10-Ks to generate a set of dynamic industry structure and competition. They find the new industry classification is not only useful to understand how industry structure changes over time but also to learn how firms react to dynamic changes within and around their product markets. Comparing to equities market, cryptocurrencies market only serves Blockchain-based start-ups with very few reports on earnings or related fundamental information. Although White Paper is required to describe company business models and future plans before Initial Coin Offering (ICO), the uncertainty remains high given fake ICOs and unpredictable market environment or regulations changes. This makes the content of ICO White Paper not as much informative as 10-Ks. Consequently, instead of the White Paper, we extract representative fundamental information of each mining contract, i.e., algorithm and proof types, as additional information input given that Blockchain technology mainly depends on its algorithm and rewarding system. In addition, we use return comovement to

proxy the fundamental similarity of each cryptocurrency to enrich the dataset. Since the cryptocurrencies are traded in high-frequency, return information is particularly important as it serves as a timely information for understanding the dynamics of market structure.

Apart from data scarcity, there are still several technical obstacles when we start dealing with the real data. Firstly, in order to build up network linkages using coin returns, for each coin we need to select from nearly 200 candidate coins whose returns are significantly correlated with it. Due to overfitting issue, simple linear regression is apparently inappropriate for this situation. Besides, we also need to incorporate fundamental characteristics into the return-inferred network to assist classifying cryptocurrencies. Therefore, to tackle these problems, we develop an efficient method to classify the cryptocurrencies into 5 groups and provide theoretical justifications to guarantee its consistency. Specifically, we first use the adaptive Lasso to recursively regress each coin's return on other coins to help us choose the crypto coins that possess the most significant explanatory power, and we take this significance as a network linkage between the coins in each period. Then, based on the dynamic degree corrected stochastic blockmodel (DDCBM), we design a dynamic covariate-assisted spectral clustering algorithm to incorporate both historical linkage information and fundamental characteristics into classification procedures.

In the empirical study, we estimate the group memberships of each cryptocurrency using the first two and a half years observations, namely, from 2015-07-01 to 2017-12-30. Then, we proceed to investigate the economic meanings as well as investment implications behind them using the most recent observations, i.e. from 2018-01-01 to 2018-03-31. By comparing the within-group centrality score with the cross-group centrality score of each group, we find our algorithm captures both fundamental characteristics and return information better than the benchmark algorithm in all cases. Scrutinizing the composition of fundamental characteristics in each group, we find the group with the most rarely used algorithm and proof types suffers strongest return reversal. Moreover, a contrarian trading strategy shows that the low centrality group gains the highest profit with a daily return of 5.01%, and this is statistically significantly higher than the daily return of high centrality group which is 1.34%.

This paper makes several important contributions to classic finance as well as FinTech

literatures. Firstly, we provide a new machinery for studying cryptocurrencies market segmentation that can be applied to a wide variety of assets. Specifically, we extend spectral clustering methods (see Binkiewicz et al., 2017; Zhang et al., 2017; Tao, 2018, etc.) to identify communities in dynamic networks in presence of both time-evolving membership and node covariates. To make a full use of relevant information, we faces challenges caused by the features of real data, namely time dependency, degree heterogeneity, sparsity and node covariates. In this case, our newly proposed the community detection method can resolve all the aforementioned data issues at one shot. In the meantime, this method can also be simply extended to cover more asset specific characteristics for a better classification purpose.

Secondly, we contribute to the existing literature on investors' behavioural bias, in most of which short-term return reversal is a robust and economically significant evidence. For instance, Jegadeesh (1990) adopts a reversal strategy that buys and sells stocks on the basis of their prior-month returns and holds them for one month, resulting in profits of about 2% per month spanning from 1934 to 1987. Two possible explanations of short-term reversal profits that are widely accepted by previous literature. In majority (see Shiller, 1984; Black, 1986; Subrahmanyam, 2005, etc.), short-term reversal profits indicates that investors overreact to information, or fads, or simply cognitive errors. Others suggest that short-term reversal profits are generated by the price pressure while the short-term demand curve of a stock is downwardly sloping and/or the supply curve is upwardly sloping (Grossman and Miller, 1988; Jegadeesh and Titman, 1995). Campbell and Wang (1993) find that uniformed trading activities trigger a temporary concession in price, which, when absorbed by those who provide liquidity, will lead to a price reversal as a compensation for the liquidity providers. In addition, Berkman et al. (2012) provide empirical evidence that attention-generating events (high absolute returns or strong net buying by retail investors) contribute to higher demand by individual investors, generating temporary price pressure at the open and thus the elevated overnight returns that are reversed during the trading day. In our paper, we also document a strong return reversal effect and provide new explanations to it through investors behaviour channel. In particular, we construct a novel measure of "valuation hardness" using the centrality scores of fundamental-inferred

network structure, which reflects the popularity of a fundamental setting employed by the cryptocurrencies market. We then suggest the hypothesis that cryptocurrencies with low centrality scores (rare common settings in the fundamental algorithm and proof types) tend to be hard-to-value cryptocurrencies due to less peer fundamental information is revealed by the market. Consistent with the spirit of Berkman et al. (2012), we find these "hard-to-value" cryptocurrencies reveal stronger return reversal effect than those easy-to-value ones. Most recently, Detzel et al. (2018) provide the first equilibrium model featuring technical traders and assets without cash flows. In particular, the paper suggests Bitcoin traders must rely heavily on the price trajectories which reflect the common belief of the investors in the market. In our case, we further point out that investors not only collect information from coin's historical price but also from its peer cryptocurrencies in terms of similar fundamental settings. In this case, cryptocurrencies that adopt unique technologies (i.e., algorithms and proof types) have less information available in the market due to fewer peer fundamental settings employed by other cryptocurrencies than those adopting common technologies. This will result in a stronger investors' behaviour bias.

Last but not the least, we deepen the understanding of cryptocurrencies market in both market segmentation and portfolio construction. Cryptocurrency is now a fast emerging alternative asset class that urges for deeper academic understanding and explorations. Numerous literature in this area study asset pricing inference from different angles while limited work shows economic linkage of cryptocurrency fundamentals and its performance. Ong et al. (2015) evaluate the potential of cryptocurrency using social media data and find that merged pull requests of GitHub, number of merges, number of active account and number of total comments are the four key variables determining the market capitalization of cryptocurrency. Elendner et al. (2015) study the top 10 cryptocurrencies by market capitalization and find that the returns are weakly correlated with each other. Trimborn and Härdle (2017b) construct CRIX, a market index which consists of a selection of cryptocurrencies that represent the whole cryptocurrencies market, and show that the cryptocurrencies market which is momentarily dominated by Bitcoin still needs a representative index since Bitcoin does not lead the market. Given the low liquidity in the current altcoin market compared to traditional assets, Trimborn and Härdle (2017a) propose a

Liquidity Bounded Risk-return Optimization (LIBRO) approach that takes into account liquidity issues by studying the Markowitz framework under the liquidity constraints. Chen et al. (2018) study the option pricing for cryptocurrency based on a stochastic volatility model with correlated jumps. Lee et al. (2017) compare cryptocurrencies with traditional asset classes and find that cryptocurrency provides additional diversification to the mainstream assets, hence improving the portfolio performance. As an innovation of FinTech, cryptocurrency fundamentals display different features from the traditional assets and these features indeed bring in certain new effects on the price evolution. In this case, our paper contributes to better understanding to the fundamental of the market structure by proposing a new clustering method. By dividing cryptocurrencies into five groups according to our classification method, we provide solid empirical evidence to how fundamental characteristics take an impact on the cryptocurrencies prices.

The remainder of the paper is organized as follows. In section 2, we introduce the model and the method designed for estimating the dynamic group structure, and we demonstrate the effectiveness of our method by simulation. In section 3, we employ our method to classify the cryptocurrencies and explain the economic interpretation behind the grouping results. Then, in section 4, we check the time series and cross-sectional return predictability and demonstrate its portfolio implications. Lastly, we conclude in section 5. All the proofs and technical details are provided in the appendix.

2 Models and Methodology

In this section, we will extend the dynamic covariate-assisted spectral clustering (CASC) algorithm (Tao, 2018) to deal with the dynamic version of uni-partite *spectral-contextualized stochastic block model* (SC-SBM) proposed by Zhang et al. (2017), which will be applied to modelling the group structure of the cryptocurrencies network in the following sections. We then provide the theoretical justification of the algorithm and conduct several simulations to show the consistency of this method.

2.1 Dynamic Network Model with Covariates

To study the block structure of dynamic network, we consider a dynamic network defined as a sequence of random undirected graphs with N nodes, $G_{N,t}$, $t = 1, \cdots, T$, on the vertex set $V_N = \{v_1, v_2, \cdots, v_N\}$ which does not change over horizons. For each period, we model the uni-partite network structure with the degree-corrected *spectral-contextualized stochastic block model* (SC-SBM) introduced by Zhang et al. (2017). Specifically, we observe adjacency matrices A_t of the graph at time instances $\varsigma_t = t/T$ where $0 < \varsigma_1 < \varsigma_2 < \cdots < \varsigma_T = 1$. The adjacency matrix A_t is generated by

$$A_t(i,j) = \begin{cases} \text{Bernoulli}(P_t(i,j)), & \text{if } i < j \\ 0, & \text{if } i = j \\ A_t(j,i), & \text{if } i > j \end{cases} \tag{1}$$

where $P_t(i,j) = \Pr(A_t(i,j) = 1)$. Basically, we assume that the probability of a connection $P_t(i,j)$ is entirely determined by the groups to which the nodes i and j belong at the moment ς_t. In particular, if $z_{i,t} = k$ and $z_{j,t} = k'$, then $P_t(i,j) = B_t(z_{i,t}, z_{j,t}) = B_t(k,k')$. In this case, for any $t = 1, \cdots, T$, one has the population adjacency matrix

$$\mathcal{A}_t := \mathbb{E}(A_t) = Z_t B_t Z_t^\top, \tag{2}$$

where $Z_t \in \{0,1\}^{N \times K}$ is the *clustering matrix* such that there is only one 1 in each row and at least one 1 in each column.

Since conventional stochastic blockmodel presumes that each node in the same group should have same expected degrees, Karrer and Newman (2011) propose a degree correction approach to overcome this unrealistic assumption. Following Karrer and Newman (2011), we introduce the *degree parameters* $\psi = (\psi_1, \cdots, \psi_N)$ to capture the degree heterogeneity of the groups. In particular, the edge probability between node i and j at time t is given by

$$P_t(i,j) = \psi_i \psi_j B_t(z_{i,t}, z_{j,t}). \tag{3}$$

In addition, to resolve identifiability issue, Karrer and Newman (2011) impose the restriction that

$$\sum_{i \in \mathcal{G}_k} \psi_i = 1, \quad \forall k \in \{1, 2, \cdots, K\}. \tag{4}$$

Then, denote Diag(ψ) by Ψ, the population adjacency matrices for dynamic degree-corrected spectral-contextualized stochastic blockmodel (SC-DCBM) is

$$\mathcal{A}_t = \Psi Z_t B_t Z_t^\top \Psi, \tag{5}$$

Define the regularized graph Laplacian as

$$L_{\tau,t} = D_{\tau,t}^{-1/2} A_t D_{\tau,t}^{-1/2}, \tag{6}$$

where $D_{\tau,t} = D_t + \tau_t I$ and D is a diagonal matrix with $D_t(i,i) = \sum_{j=1}^{N} A_t(i,j)$. As shown in Chaudhuri et al. (2012), including the regularization parameter can help improve the spectral clustering performance on sparse networks. Therefore, we choose the regularized parameter τ_t according to the convention (Qin and Rohe, 2013) by taking the value of average node degree in each period, i.e. $\tau_t = N^{-1} \sum_{i=1}^{N} D_t(i,i)$.

Now, we introduce the bounded covariates associated to each vertex $X(i) \in [-J, J]^R$, $i = 1, \cdots, N$ for all $t = 1, \cdots, T$. In Binkiewicz et al. (2017), they add the covariance $XX^\top$ to the regularized graph Laplacian and perform the spectral clustering on the *similarity matrix*, and Tao (2018) extends the static similarity matrix to cover the dynamic case as below:

$$S_t = L_{\tau,t} + \alpha_t C. \tag{7}$$

where $C = XX^\top$ and $\alpha_t \in [0, \infty)$ is a tuning parameter that controls the informational balance between $L_{\tau,t}$ and X in the leading eigenspace of S_t. As a generalization of the model, Zhang et al. (2017) refines Binkiewicz et al. (2017) by replacing C with $C_w = XWX^\top$. Similarly, we make the same generalization to Tao (2018) by substituting C with the new covariate assisted component $C_t^w = XW_tX^\top$, and the population similarity matrix now becomes

$$\mathcal{S}_t = \mathcal{L}_{\tau,t} + \alpha_t \mathcal{C}_t^w, \tag{8}$$

where $\mathcal{L}_{\tau,t} = \mathcal{D}_{\tau,t}^{-1/2} \mathcal{A}_t \mathcal{D}_{\tau,t}^{-1/2}$ and $\mathcal{C}_t^w = \mathcal{X}\mathcal{W}_t\mathcal{X}$.

This is a non-trivial generalization as it addresses several limitations of dynamic CASC (Tao, 2018). Firstly, W_t creates a time-varying interaction between different covariates. For instance, we may think of different refined algorithms that stem from the same origins. Such inheritance relationships will potentially leads to an interaction between the cryptocurrencies. In addition, as time goes by, some algorithms may become more and more

popular while the others may near extinction. Thus, this interaction would also change over time. These interactions are not included in C.

Secondly, we can easily select covariates by setting certain elements of W_t to zero. This is necessary as it helps us to model the evolution of technologies. At some point of time, some cryptographic technology may be eliminated due to upgrading or cracking. Therefore, W_t offers us the flexibility to exclude covariates which cannot be easily done with C.

Lastly, as suggested in Zhang et al. (2017), C presumes that similarity in covariates leads to high probability of node connection. However, it may not be true in cryptocurrencies network. Due to the open source nature of blockchain, cryptocurrency developers can easily copy and paste the source codes and launch a new coin without any costs. Consequently, it causes severe homogeneity in cryptocurrencies market. Nevertheless, this homogeneity does not necessarily end up with co-movement of prices in reality. Some of coins are even negatively correlated with each other. In this case, we can just set $W_t(i,i)$ to be negative and C_t^w will in the end bring the coins with different technologies closer in the similarity matrix.

2.2 Dynamic Covariate-assisted Spectral Clustering

To perform dynamic CASC, we face two major difficulties: (i) setting of W_t, (ii) estimation the similarity matrix with dynamic network information. For the first issue, we follow Zhang et al. (2017) by setting $W_t = X^\top L_{\tau,t} X$ which measures the correlation between covariates along the graph. For the second issue, we follow Pensky and Zhang (2017) by constructing the estimator of $\mathcal{S}_t$ with discrete kernel method to weigh the historical network information. Specifically, we first pick an integer $r \geq 0$, and obtain three pairs of sets of integers

$$\mathcal{F}_r = \{-r, \cdots, 0\}, \quad \mathcal{D}_r = \{T - r + 1, \cdots, T\},$$

and we assume that $|W_{r,l}(i)| \leq W_{\max}$, where $W_{\max}$ is independent of r and i, and satisfies

$$\frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} i^k W_{r,l}(i) = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{if } k = 1, 2, \cdots, l. \end{cases} \tag{9}$$

Obviously, the $W_{r,l}$ is a discretized version of continuous boundary kernel that only

weighs the historical observations. This kernel assigns the more recent similarity matrices with the higher scores and . To choose an optimal bandwidth r, Pensky and Zhang (2017) propose an adaptive estimation procedure using Lepski's method. Here, we directly apply their results and construct the estimator for edge connection matrices S_t as below

$$\widehat{\mathcal{S}}_{t,r} = \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} W_{r,l}(i) S_{t+i}. \tag{10}$$

Once we obtain the estimate of $\mathcal{S}_t$, we then consider the eigen-decomposition of $S_t = U_t \Lambda_t U_t^\top$ for each $t = 1, 2, \cdots, T$. As discussed in Lei and Rinaldo (2015), the matrix U_t now may have more than K distinct rows as a result of degree correction whereas the rows of U_t still only point to at most K directions. Therefore, we apply the spherical clustering algorithm to find a cluster structure among the rows of a normalized matrix U_t^+ with $U_t^+(i, *) = U_t(i, *)/\|U_t(i, *)\|$. Specifically, we consider the following spherical k-means spectral clustering:

$$\left\| \widehat{Z}_t^+ \widehat{Y}_t - \widehat{U}_t^+ \right\|_F^2 \leq (1 + \varepsilon) \min_{\substack{Z_t^+ \in \mathcal{M}_{N_+, K} \\ Y_t \in \mathbb{R}^{K \times K}}} \left\| Z_t^+ Y_t - \widehat{U}_t^+ \right\|_F^2 \tag{11}$$

Finally, we extend $\widehat{Z}_t^+$ to obtain $\widehat{Z}_t$ by adding $N - N_+$ many canonical unit row vectors at the end. $\widehat{Z}_t$ is the estimate of Z_t from this method. The detailed algorithm is summarized as below.

To ensure the performance of the dynamic covariate-assisted spectral clustering method, we first make some assumptions on the graph that generates the dynamic network. The major assumption we need here is the *assortativity* which ensures the nodes within the same cluster are more likely to share an edge than nodes in two different clusters.

Assumption 1. *The dynamic network is composed of a series of assortative graphs that are generated under the stochastic block model with covariates whose block probability matrix B_t is positive definite for all $t = 1, \cdots, T$.*

Assumption 2. *There are at most $s < \infty$ number of nodes can switch their memberships between any consecutive time instances.*

Assumption 3. *For $1 \leq k \leq k' \leq K$, there exists a function $f(\cdot; k, k')$ such that $B_t(k, k') = f(\varsigma_t; k, k')$ and $f(\cdot; k, k') \in \Sigma(\beta, L)$, where $\Sigma(\beta, L)$ is a Hölder class of functions $f(\cdot)$ on*

Algorithm 1: Covariate-Assisted Spectral Clustering in the Dynamic SC-DCBM

Input : Adjacency matrices A_t for $t = 1, \cdots, T$;

Covariates matrix X;

Number of communities K;

Approximation parameter ε.

Output: Membership matrices Z_t for any $t = 1, \cdots, T$.

1 Calculate regularized graph Laplacian $L_{\tau,t}$ and weight matrix W_t.

2 Estimate $\mathcal{S}_t$ by $\widehat{\mathcal{S}}_{t,r}$ defined in (10).

3 Let $\widehat{U}_t \in \mathbb{R}^{N \times K}$ be a matrix representing the first K eigenvectors of $\widehat{\mathcal{S}}_{t,r}$.

4 Let N_+ be the number of nonzero rows of $\widehat{U}_t$, then obtain $\widehat{U}^+ \in \mathbb{R}^{N_+ \times K}$ consisting of normalized nonzero rows of $\widehat{U}_t$, i.e. $\widehat{U}_t^+(i,*) = \widehat{U}_t(i,*) / \left\|\widehat{U}_t(i,*)\right\|$ for i such that $\left\|\widehat{U}_t(i,*)\right\| > 0$.

5 Apply the $(1+\varepsilon)$-approximate k-means algorithm to the row vectors of $\widehat{U}_t^+$ to obtain $\widehat{Z}_t^+ \in \mathcal{M}_{N_+,K}$.

6 Extend $\widehat{Z}_t^+$ to obtain $\widehat{Z}_t$ by arbitrarily adding $N - N_+$ many canonical unit row vectors at the end, such as, $\widehat{Z}_t(i) = (1, 0, \cdots, 0)$ for i such that $\left\|\widehat{U}_t(i,*)\right\| = 0$.

7 Output $\widehat{Z}_t$.

$[0,1]$ *such that* $f(\cdot)$ *are* ℓ *times differentiable and*

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L|x - x'|^{\beta-\ell}, \text{ for any } x, x' \in [0,1], \tag{12}$$

with ℓ *being the largest integer smaller than* β.

Assumption 4. *Let* $\lambda_{1,t} \geq \lambda_{2,t} \geq \cdots \geq \lambda_{K,t} > 0$ *be the* K *largest eigenvalues of* $\mathcal{S}_t$ *for each* $t = 1, \cdots, T$. *In addition, assume that*

$$\underline{\delta} = \inf_t \{\min_i \mathcal{D}_{\tau,t}(i,i)\} > 3\ln(8NT/\epsilon) \quad \text{and} \quad \alpha_{\max} = \sup_t \alpha_t \leq \frac{a}{NRJ^2\xi},$$

with

$$a = \frac{3\ln(8NT/\epsilon)}{\underline{\delta}} \quad \text{and} \quad \xi = \max(\sigma^2\|L_\tau\|_F\sqrt{\ln(TR)}, \sigma^2\|L_\tau\|\ln(TR), NRJ^2/\underline{\delta}),$$

where $\sigma = \max_{i,j}\|X_{ij} - \mathcal{X}_{ij}\|_{\phi_2}$, $L_\tau = \sup_t L_{\tau,t}$.

To establish the consistency of covariate-assisted spectral clustering for dynamic SBM, we need to figure out the upper bounds for the misclustering rates. Following Binkiewicz et al. (2017), we denote $C_{i,t}$ and $\mathcal{C}_{i,t}$ as the cluster centroids of the ith node at time t generated using k-means clustering on U_t and $\mathcal{U}_t$ respectively. Then, we define the set of misclustered nodes at each period to be

$$\mathbb{M}_t = \left\{i\text{: } \left\|C_{i,t}\mathcal{O}_t^\top - \mathcal{C}_{i,t}\right\| > \left\|C_{i,t}\mathcal{O}_t^\top - \mathcal{C}_{j,t}\right\|, \text{for any } j \neq i\right\}, \tag{13}$$

where $\mathcal{O}_t$ is a rotation matrix that minimizes $\|U_t\mathcal{O}_t^\top - \mathcal{U}_t\|_F$ for each $t = 1, \cdots, T$.

The error has two folds. The first source of the error is the estimation error of $\mathcal{S}_t$ using the discrete kernel estimator. The second source of the clustering error comes from the spectral clustering algorithm. Then, we can derive the uniform upper bound for the misclustering rate for the covariate-assisted spectral clustering for dynamic SC-DCBM.

Theorem 1. *Let clustering be carried out according to the Algorithm 1 on the basis of an estimator* $\widehat{\mathcal{S}}_{t,r}$ *of* $\mathcal{S}_t$. *Let* $Z_t \in \mathcal{M}_{N,K}$ *and* $P_{\max} = \max_{i,t}(Z_t^\top Z_t)_{ii}$ *denote the size of the largest block over the horizons. Then, under Assumption 1-4, as* $N, T, R \to \infty$ *with* $R = o(N)$, *the misclustering rate satisfies*

$$\sup_t \frac{|\mathbb{M}_t|}{N} \leq \frac{c(\varepsilon)KW_{\max}^2}{m_z^2 N\lambda_{K,\max}^2}\left\{(4+2c_w)\frac{b}{\underline{\delta}^{1/2}} + \frac{2K}{b}(\sqrt{2P_{\max}rs} + 2P_{\max}) + \frac{NL}{b^2 \cdot l!}\left(\frac{r}{T}\right)^\beta\right\}^2.$$

with probability at least $1-\epsilon$*, where* $\lambda_{K,\max} = \max_t\{\lambda_{K,t}\}$ *with* $\lambda_{K,t}$ *being the* K*th largest absolute eigenvalue of* $\mathcal{S}_t$*, where* $b = \sqrt{3\ln(8NT/\epsilon)}$*,* $\lambda_{K,\max} = \max_t\{\lambda_{K,t}\}$ *and* $c(\varepsilon) = 2^9(2+\varepsilon)^2$*.*

The last problem we face is the choice of tuning parameter, r and α, and the estimation of number of groups, K. For the choice of r, we directly apply Lemma 4.5 of Tao (2018) and Lepski's method, and obtain

$$\widehat{r} = \max\left\{0 \le r \le T/2 : \left\|\widehat{\mathcal{S}}_{t,r} - \widehat{\mathcal{S}}_{t,\rho}\right\| \le 4W_{\max}\sqrt{\frac{N\|\mathcal{S}_t\|_\infty}{\rho \vee 1}},\ \text{for any } \rho < r\right\}. \tag{14}$$

Next, for choice of α_t, following Tao (2018), we just choose α_t to achieve the balance between $L_{\tau,t}$ and C_t^w, i.e,

$$\alpha_t = \frac{\lambda_K(L_{\tau,t}) - \lambda_{K+1}(L_{\tau,t})}{\lambda_1(C_t^w)}. \tag{15}$$

Lastly, for the estimation of K, we have several choices. Wang and Bickel (2017) propose a pseudo likelihood approach for choosing the number of clusters for stochastic blockmodel without covariates and prove its consistency. Chen and Lei (2017) propose a network cross-validation procedure to estimate the number of clusters by utilizing the adjacency information of stochastic blockmodel. Most recently, Li et al. (2016) refines the network cross-validation approach by proposing an edge sampling algorithm. In our case, we can directly apply network cross-validation approach by inputting the similarity matrix instead of adjacency matrix[1].

2.3 Monte Carlo Simulations

In this section, we carry out some simulations under different model setups and make comparisons with existing clustering methodologies to demonstrate the finite sample performance of our clustering algorithms. Our benchmark algorithms are the dynamic degree corrected spectral clustering for sum of squared adjacency matrix (DSC-DC) by Bhattacharyya and Chatterjee (2017) on the averaged similarity matrix and the dynamic spectral clustering method (DSC-PZ) by Pensky and Zhang (2017).

[1] As we will show in the subsequent section, when we use dummy variables to indicate different technology attributes, the covariate matrix C_t^w behaves just like an adjacency matrix. Therefore, we can directly apply network cross-validation to similarity matrix in our study.

Firstly, we set the block probability matrix B_t to be

$$B_t = \frac{t}{T}\begin{bmatrix} 0.9 & 0.6 & 0.3 \\ 0.6 & 0.3 & 0.4 \\ 0.3 & 0.4 & 0.8 \end{bmatrix}, \text{ with } 1 \leq t \leq T.$$

and the order of polynomials for kernel construction $L = 4$ for all simulations. The number of communities K is assumed to be known throughout the simulations, and the time-invariant node covariates is set to $R = \lfloor \ln(N) \rfloor$ dimensional with values $X(i,j) \overset{i.i.d}{\sim}$ Uniform(0,10) with $i \in \{1, \cdots, N\}$ and $j \in \{1, \cdots, R\}$. All experiments are replicated 100 times

Our first simulation checks the clustering performance under growing size of the network. The number of nodes in the network varies from 10 to 100 with step size 5. The time span is $T = 10$. The results are summarized in Figure 1(a). Clearly, we can observe that, as the size of the graph grows, the misclustering rates of all spectral clustering methods based on literature decrease sharply and our method dominates DSC-PZ. Besides, it can be observed that DSC-DC perform very badly when the size of the network is small (below 50) while CASC-DC still possesses an acceptable misclustering rate. This shows clearly that our method gains a huge advantage over the existing spectral clustering method for dynamic SC-DCBMs. In addition, it shows that although using covariate alone (DSC-Cw) for clustering is unsatisfactory, we can still add covariate to adjacency matrix for clustering and achieve a better grouping result. This evidence to some extent justifies the use of covariate information in our methodology.

Next, we check the performance of our method comparing with other methods under growing maximal number of group membership changes. In the simulation settings, we hold the total number of vertices to be 100, then we allow the each-period group membership changes, s, to vary in set $\{0, N/50, N/25, N/20, N/10, N/5, N/4, N/2, N\}$. The total number of horizons is $T = 10$ and the results are summarized in Figure 1(b). As shown in the figure, we can conclude that all methods are sensitive to total number of group membership changes. In other words, the more unstable the group membership is, the higher the misclustering rate will be. In spite of that, our method still achieves a lower misclustering rate than the benchmark methods in all cases.

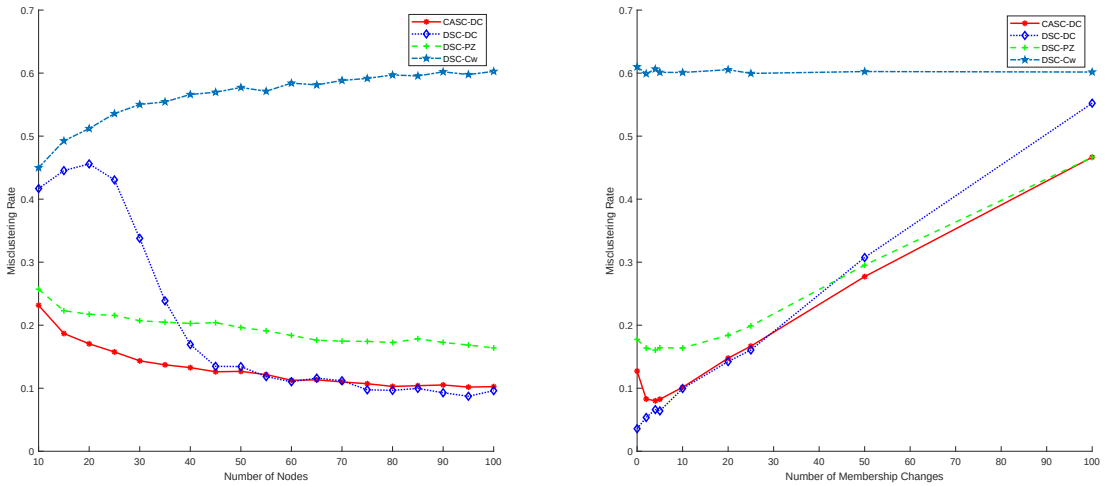


Figure 1: This figure reports the misclustering rate of different spectral clustering algorithms. CASC-DC stands for the covariate-assisted spectral clustering method we developed for dynamic degree corrected stochastic blockmodel. DSC-DC denotes the dynamic spectral clustering methods developed in Bhattacharyya and Chatterjee (2017). DSC-PZ denotes the dynamic spectral clustering methods developed in Pensky and Zhang (2017). DSC-Cw denotes the dynamic spectral clustering only based on covariate information. In subfigure (a), the number of nodes varies from 10 to 100, and the number of membership changes is fixed as $s = N^{1/2}$. In subfigure (b), the number of nodes is fixed as 100, while the number of membership changes varies from 0 to 100. In both figures, the horizon $T = 10$ and all simulations are repeated 100 times.

3 Network Construction

In this section, we study the latent group structure of cryptocurrencies market by applying our covariate-assisted spectral clustering methods to analysing the dynamic networks formulated by cryptocurrencies returns. We first introduce how we identify linkages between cryptocurrencies using adaptive Lasso regression, and construct the network. Then, we try to answer the question how do fundamental information and return structure jointly determine the cryptocurrencies market segmentation.

3.1 Data and Variables

We collect data on the daily historical price, trading volume and contract information of cryptocurrency from the website *Cryptocompare.com*, which is an interactive platform and provides us a free API access to download the data. We start with the top 200 cryptocurrencies[2] for our analysis, and the number then becomes 199 after excluding those with incomplete contract information data. The whole sample period spans from 2015-08-01 to 2018-03-31 with in-sample period for community detection from 2015-07-01 to 2017-12-31 and the rest 3 months as out-of-sample period (2018-01-01 to 2018-03-31).

For node covariates, which are fixed over time, we collect algorithm and proof types from the contract information. In fact, these covariates are not chosen arbitrarily. Instead, we have profound reasons for selecting these characteristics.

Algorithm, which is in short for *hashing algorithm*, plays a central role in determining the security of the cryptocurrencies. For each cryptocurrency, there is a hash function in mining cryptography, e.g. Bitcoin uses double SHA-256 and Litcoin uses Scrypt. As security is one of the most important features of cryptocurrencies, the hash algorithm naturally determines the intrinsic value of a cryptocurrency. In the above example, scrypt system was put into use with cryptocurrencies in an effort to improve upon the SHA256 protocol which preceded it and which bitcoin is based on. Specifically, scrypt was employed as a solution to prevent specialized hardware from brute-force efforts to out-mine others for bitcoins. As a result, Scrypt altcoins require more computing effort per unit, on average,

[2]We sort all cryptocurrencies according to the history, trading volume and maximum daily transaction price and pick up the top 200 cryptocurrencies as of 2017-12-31.

than the equivalent coin using SHA256. The relative difficulty of the algorithm confers relative value.

Proof Types, or proof system/protocol, is an economic measure to deter denial of service attacks and other service abuses such as spam on a network by requiring some work from the service requester, usually meaning processing time by a computer. For each cryptocurrency, it will at least choose one of the protocol as transaction verification method, e.g. Bitcoin and Ethereum use Proof-of-Work, Diamond and Blackcoin use Proof-of-Stake. In this case, how efficient of the proof protocol determines the reliability, security and effectiveness of the coin transactions, which will also affect the value of the cryptocoins.

3.2 Return Inferred Network Structure

As introduced in the previous section, we have collected a data sample of 199 cryptocurrencies. Therefore, to improve estimation and inference as well as to avoid over-fitting for the general regression, we employ the adaptive Lasso proposed by Zou (2006). The adaptive Lasso estimates are defined as:

$$\hat{b}_i^* = \arg\min \left\| r_{i,t}^s - \alpha_i - \sum_{j=1}^{N} b_{i,j} r_{j,t} \right\|^2 + \lambda_i \sum_{j=1}^{N} \hat{w}_i |b_{i,j}|, \tag{16}$$

where $r_{j,t}$ is the standardized return for cryptocurrency j, $\hat{b}_i^* = (\hat{b}_{i,1}^*, \cdots, \hat{b}_{i,N}^*)'$ is an N dimensional vector of adaptive Lasso estimates, λ_i is a non-negative regularization parameters, and $\hat{b}_{i,j}$ is the weight corresponding to $|b_{i,j}|$ for $j = 1, \cdots, N$ in the penalty term. Adaptive Lasso uses ℓ_1-norm penalty to shrink the parameter estimates to prevent over-fitting and hence, selecting most informative connections. Those cryptocurrencies selected by adaptive Lasso will be labelled as linked coins to the cryptocurrency i. For model estimation, we require at least 60 daily observations for each coin and set the initial estimation window as 60 days (2-month observations). We repeat this process for each cryptocurrency in each period, and finally obtain the adjacency matrix, $\mathcal{A}_t$, which can be used to form a series of undirected graphs.

Based on the adjacency matrix, we further explore the relative importance of each node by deriving the centrality of a cryptocurrency. We compute eigenvector centrality score of

cryptocurrencies, c_t, using the definition

$$\mathcal{A}_t \mathbf{c}_t = \lambda_{\max} \mathbf{c}_t, \text{ for each } t = 1, 2, \cdots, T,$$

where $\mathbf{c}_t = (c_{1,t}, c_{2,t}, \cdots, c_{N,t})'$.

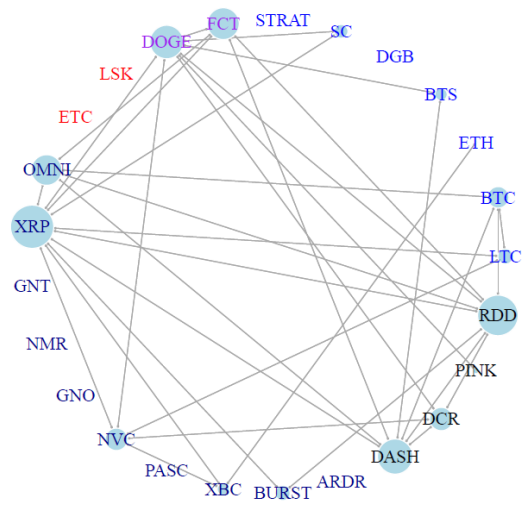
In Figure 2, we visualize some subgraphs on selected dates to illustrate the structural features of this return inferred network. Without loss of generality, we select top 5 cryptocurrencies in terms of market capitalization as of 2017-12-31 from final grouping results based on our dynamic covariate-assisted spectral clustering method. We then plot the subnetwork induced by the submatrix of the adjacency matrix on selected cryptocurrencies and dates. The colour of node labels stand for grouping results based on return inferred network structure using Bhattacharyya and Chatterjee (2017), where the group membership is fixed over time, and the node size denotes its eigenvector centrality.

Obviously, return inferred network structure is time varying and hence provides us a dynamic network structure for clustering analysis. Comparing to the fixed node features, time varying network structure delivers valuable information about investors' opinion changes. However, the return inferred network structure is not very stable over time[3] and it could be very sparse on some days, e.g. 10/25 cryptocurrencies do not have any connections to any cryptocurrencies on 2016-03-15, which would lead to inconsistent classification results. In this case, our covariate-assisted spectral clustering method comes in to solve this problem by integrating node features to assist clustering analysis.

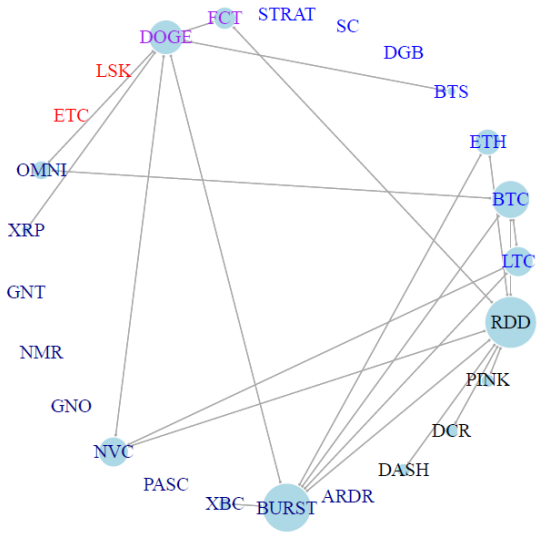
3.3 Contract Information Inferred Network Structure

To demonstrate how contract information can assist cryptocurrency classifications, we construct a contract information inferred network to illustrate how its structure differs from the return inferred network structure. We define that two cryptocurrencies are connected as long as two cryptocurrencies share at least one same fundamental characteristic. Taking Ethereum and Ethereum Classic as an example, since both of them use Ethash as their hash algorithm, these two cryptocurrencies are regarded as connected by definition. Apart from the algorithm, we also adopt proof types as additional fundamental information to

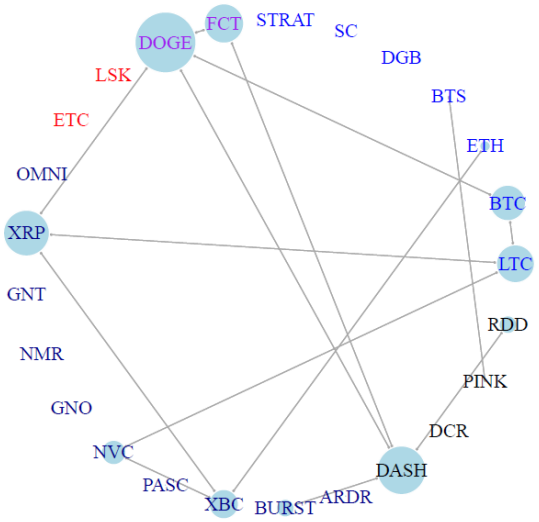
[3]This result makes sense as investors update their beliefs on daily frequency.



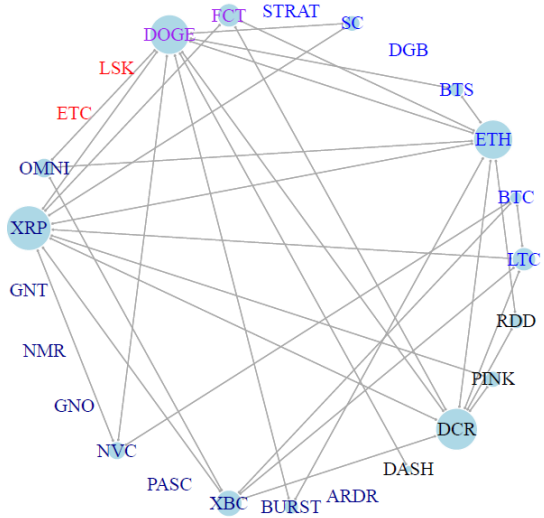
(a) 2016-03-01



(b) 2016-03-05



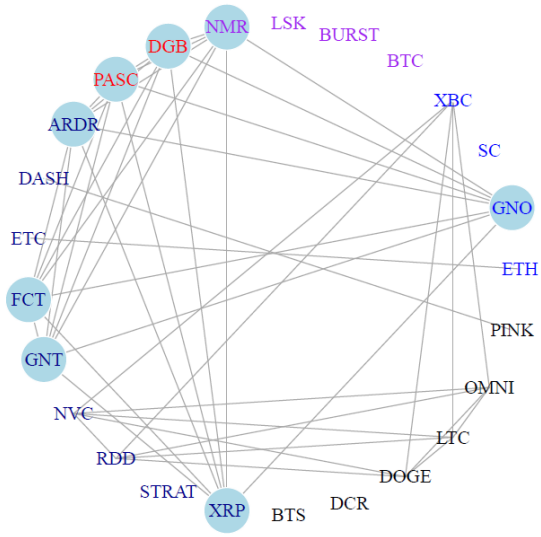
(c) 2016-03-15



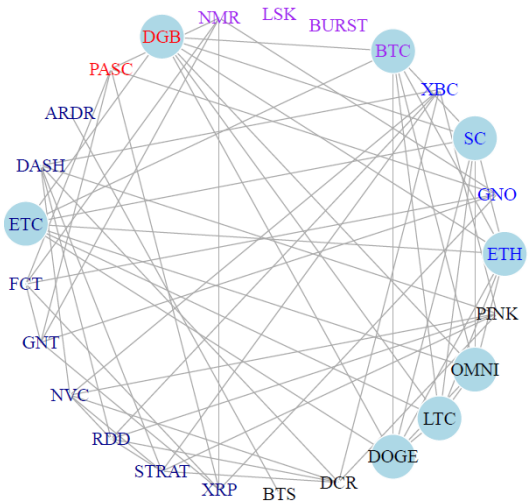
(d) 2016-03-31

Figure 2: This figure depicts the time varying of return inferred network structure. In the layout, we plot 25 cryptocurrencies, including BTC, ETH, LTC and other top cryptocurrencies within each group according combined information in terms of market capitalization as of 2017-12-31. Connection is defined from a return regression model: $r_{i,t} = \alpha_i + \sum_{j=1,j\neq i}^{N-1} b_{i,j} r_{j,t} + \epsilon_{i,t}$, where $r_{i,t}$ is the daily return on cryptocurrency i, N is the total number of cryptocurrencies. Adaptive Lasso is employed to estimate above regression and only those cryptocurrency that are being selected by adaptive lasso will be linked to cryptocurrency i. The colour of node labels stand for grouping results based on return inferred network structure using Bhattacharyya and Chatterjee (2017) and the node size denotes eigenvector centrality of a cryptocurrency.

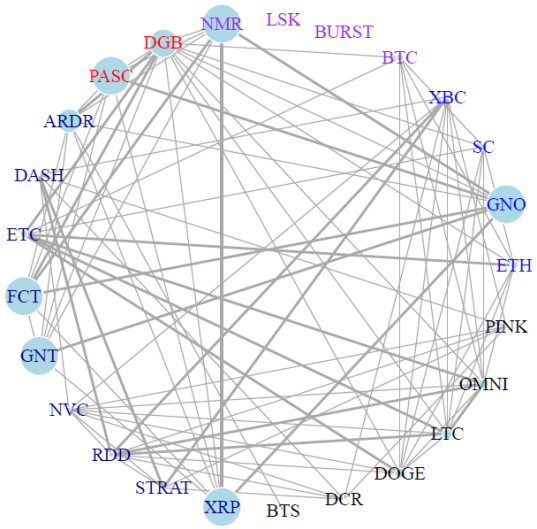
define the connections. In Figure 3, we visualize the contract information inferred network in Figure 3 using the same set of the cryptocurrencies in return inferred network.



(a) Algorithm



(b) Proof Types



(c) Combined Fundamental

Figure 3: This figure depicts the fundamental information inferred network structure. We define the connection as long as two cryptocurrencies sharing the same fundamental technology. We consider two fundamental variables, namely algorithm and proof types with their aggregated information. Node size denotes eigenvector centrality of a cryptocurrency.

As shown in Figure 3, the contract information inferred networks are less sparse than

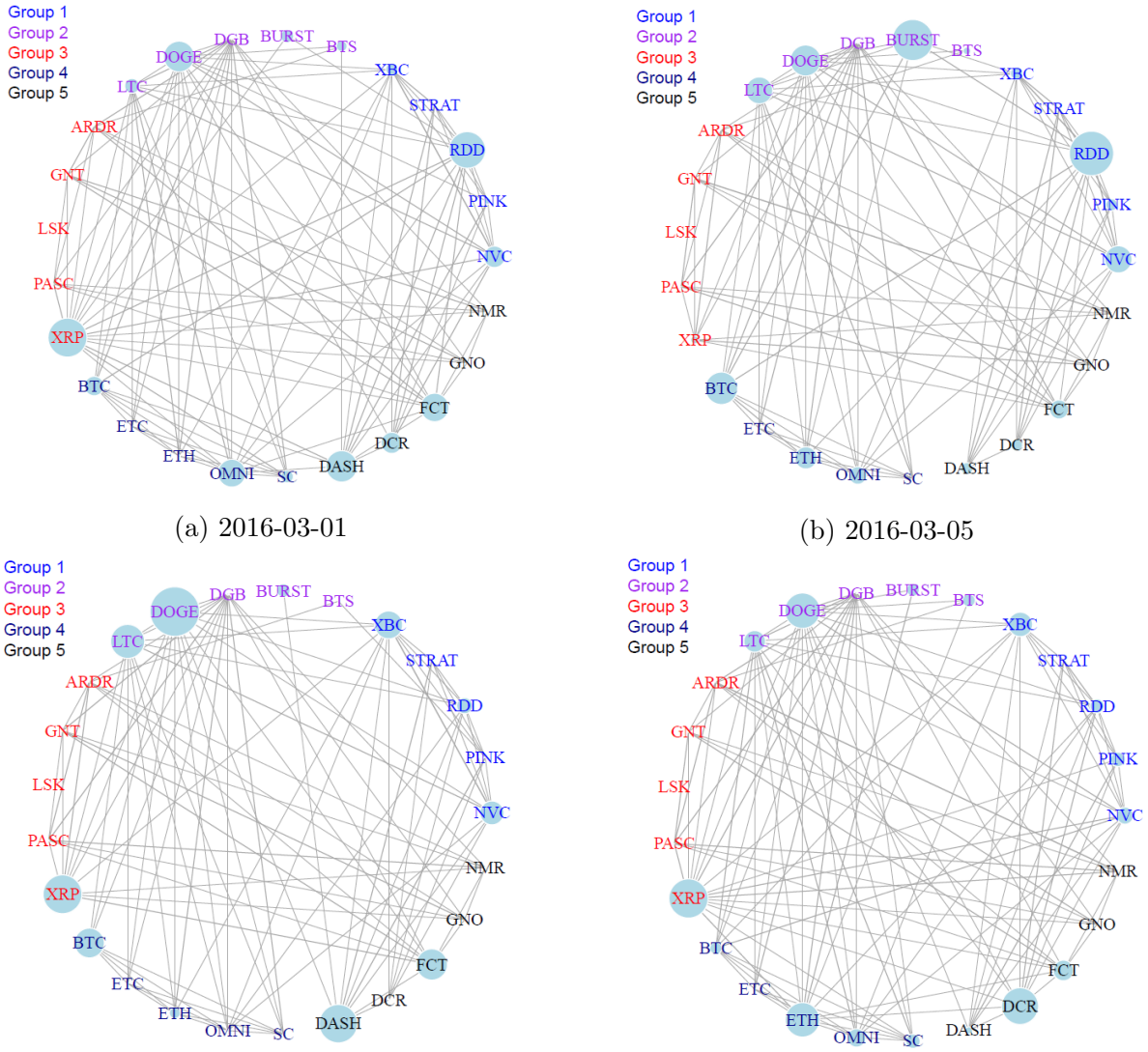
the return inferred networks. In fact, due to limited choices of algorithms and other attributions, the coins are more likely to connect with each other when using the characteristics to build up the linkages. However, it does not mean that using contract information alone to define the group structure is enough. Firstly, only relying on contract information to classify the cryptocurrencies ignores the information about time-varying connections from market investors which is particularly important for the cryptocurrencies market. Secondly, there are some difficulties in pricing those fundamental characteristics. Unlike corporate fundamentals that is straightforward to pricing equities, the relationship between the value of a cryptocurrency and its fundamental characteristics seems much more complicated. It is possible that a new algorithm does not add any valuable features to the existing algorithms. In fact, many developers simply copy and paste the blockchain source code with minor modifications on the parameters to launch a new coin for speculation purpose through ICO (Initial Coin Offering). Even though these altcoins may show little differences between their fundamental characteristics, their abilities to generate future cash flows are quite different. A good example is IXCoin, which is the first *clonecoin* of Bitcoin. Despite Bitcoin is regarded as the most successful cryptocurrency, IXCoin is not able to duplicate its success. The developer team stops working on IXCoin for months after the ICO. Reflected by its return performance, it suggests a higher risk than Bitcoin. In fact, more evidences can be found from deadcoins[4]. In summary, combining contract information with return information is necessary for revealing more informative connections between the cryptocurrencies.

3.4 Combined Network Structure

Based on the reasoning in the previous sections, we then combine the return inferred network and the contract information inferred network using similarity matrix, and we plot the combined networks in Figure 4 on selected dates as in previous sections. As shown in the figure, the linkages between cryptocurrencies not only exist within the group members, but also exist across the groups. For example, ETH is both connected to its group members, such as ETC and BTC, and is also connected to cryptocurrencies from other groups, such

[4]Matic Jurglič provides a list of deadcoins on the website: http://deadcoins.com/

as XBC and RDD in group 1 and Doge and LTC in group 2. This suggests a possible change of grouping membership in this developing market. In this paper, we provide the best fit of market structure using available data sample, and hopefully, our method will still be applicable when the market becomes more mature.



(a) 2016-03-01

(b) 2016-03-05

(c) 2016-03-15

(d) 2016-03-31

Figure 4: This figure depicts the time varying of a combined network structure based on the similarity matrix, which combines return information and contract information simutaneously. The color of node labels stand for grouping results based on combined information set using degree corrected covariate-assisted spectral clustering method and the node size denotes degree centrality of a cryptocurrency.

Comparing with the network using single information set, Figure 4 shows that combined network is denser and assigns the centrality scores to each cryptocurrency more evenly. It is interesting that we may find the cryptocurrencies who have more return linkages will have less fundamental linkages. In this case, the similarity matrix balances the return and fundamental information and leads to balanced centrality scores. A further examination shows that the cryptocurrencies who gain more return linkages are likely to adopt more original algorithms or proof types, e.g. BTC and ETH. Then, from Figure 3 we can observe that those original fundamental technologies attract less audience (smaller centrality scores) comparing to the new ones. This reflects fundamental information are dominated by new technologies in the market. As a result, we would expect the grouping results will shed some light upon how technology evolutions affect the returns of cryptocurrencies.

4 Fundamental Centrality and Return Reversal

In this section, we mainly explains the economic meanings and the asset pricing implication of the grouping results. In the first place, we show our proposed clustering approach can fully capture both return and fundamental information by comparing the within-group centrality scores with the cross-group centrality scores based on the adjacencies of single information set. Then

4.1 Communities in Cryptocurrencies Network

Following the combined network structure and applying the covariate-assisted spectral clustering method, the 200 cryptocurrencies are classified into five groups, and the grouping results are summarized in Table 1. The table indicates that the largest top 5 cryptocurrencies (BTC, ETH, XRP, LTC and DASH) in terms of market cap are not necessarily categorized into the same group. For example, LTC and BTC, although the return inferred network structure suggets a good connection between these two coins, their fundamental setting is different. BTC employes SHA256 which now becomes a minority algorithm while LTC uses Script, which seems to be the second most popular algorithm in the market. Similarly, Ripple employs Multiple algorithm, which is the most popular algorithm in the

market so Ripple is different from both BTC and LTC. Its group members also tend to employ Multiple algorithm, such as Thether and Golem.

Table 1: Representative Cryptocurrencies of Each Group.

This table lists top 10 cryptocurrencies under each group by applying covariate-assisted spectral clustering to top 200 Cryptocurrencies. The estimation is based on the sample period from 2015-08-01 to 2017-12-31.

Group ID	**Cryptocurrency**
Group 1	Stratis, PIVX, BitcoinDark, ReddCoin, FairCoin, BlackCoin, NAV Coin, Novacoin, Energycoin
Group 2	Litecoin, BitShares, Dogecoin, DigiByte, Nxt, SysCoin, MonaCoin, Gulden, PotCoin
Group 3	Ripple, Tether, Veritaseum, Waves, Iconomi, Lisk, Golem, Augur, Stellar Lumens, Status
Group 4	Bitcoin, Ethereum, Ethereum Classic, Monero, Zcash, Steem, Siacoin, GameCredits, Nexus, Ubiq
Group 5	Dash, Gnosis, Factom, Decred, Numeraire, Etheroll, Blocknet, Namecoin, CloakCoin, BitBay

To further demonstrate how reasonable our classification results are, we compare with our benchmark method introduced in Bhattacharyya and Chatterjee (2017) by checking the differences between within-group connections and cross-group connections. In Bhattacharyya and Chatterjee (2017), they develop the spectral clustering method for a dynamic stochastic blockmodel with time-varying block probability and fixed group membership in the absence of node covariates. We admit this could be an unfair comparison since our method has taken into account for more information and studied a much more complicated model. However, this is the only spectral clustering method available for the dynamic

stochastic block in the literature by far. To avoid data mining issue, we add on another contract information, maximum coin supply, which is not controled in our estimation process as an additional test. Intuitively, if the grouping method fully captures the relevant information, within-group connections should be stronger than the cross-group connections, in other word, the difference between them should be positive. The within-group connections and cross-group connections are defined as below:

$$\textit{Within-Group Connection}_i = \frac{\#\text{ of Degrees of Coins within Group } i}{4N_i},$$

$$\textit{Cross-Group Connection}_i = \frac{\#\text{ of Degrees of Coins between Group } i \text{ and other Groups}}{4\bar{N}_i},$$

Table 2 summarizes within-group connections and cross-group connections of different information set, including both returns and the contract information. Panel A reports the average return inferred connections over the sampling periods. The difference between mean of within-group connection and cross-group connection is calculated with corresponding significance level. Panel B and Panel C report algorithm inferred connections and proof types inferred connections respectively, which are constants over time. The differences between within-group connection and cross-group connection are reported in Table 2 as below.

According to Panel A in Table 2, the return information are well captured in Bhattacharyya and Chatterjee (2017)'s model as for majority of the groups, the within-group connections are significantly higher than the cross-group connections. For example, the full sample within-group connection is 0.104, which is higher than the cross-group connections by 0.005. However, for contract information (Panel B and C), the results become much weak. Both types of contract information suggest the majority groups have cross-group connections more than within-group connections, indicating that the benchmark model cannot accommodate the contract information to a large extent.

On the contrary, in results of Table 3 obtained through our covariate-assisted spectral clustering method, within-group connections are much stronger than cross-group connections for both return and contract information set. As expected, our method captures return information much better than the benchmark model in terms of the magnitude of difference between within- and cross-group connections. In addition, our method can better detect fundamental grouping information, namely, the overall difference between the

Table 2: Within-group Connection and Cross-group Connections by Bhattacharyya and Chatterjee (2017)

This table reports within-group connection and cross-group connections based on Bhattacharyya and Chatterjee (2017). Panel A reports average return inferred connections across sample period. Panel B and Panel C report algorithm inferred connections and proof types inferred connections respectively. Connections are defined as

$$\textit{Within-Group Connection}_i = \frac{\#\text{ of Degrees of Coins within Group } i}{4N_i},$$
$$\textit{Cross-Group Connection}_i = \frac{\#\text{ of Degrees of Coins between Group } i \text{ and other Groups}}{4\bar{N}_i}.$$

*, **, and *** indicates statistical significance at the 10%, 5% and 1% levels respectively.

	Return			Algorithm			Proof Types		
	Within	Cross	Diff.	Within	Cross	Diff.	Within	Cross	Diff.
G1	0.110	0.106	0.004***	0.186	0.197	-0.012	0.294	0.248	0.046
G2	0.100	0.097	0.003	0.174	0.200	-0.027	0.236	0.242	-0.006
G3	0.118	0.107	0.010***	0.287	0.238	0.050	0.177	0.220	-0.044
G4	0.111	0.092	0.019***	0.222	0.213	0.009	0.231	0.236	-0.005
G5	0.082	0.093	-0.012***	0.186	0.196	-0.010	0.241	0.235	0.006
All	0.104	0.099	0.005***	0.211	0.209	0.002	0.236	0.236	0.000

Table 3: Within-group Connection and Cross-group Connections by Dynamic CASC.

This table reports within-group connection and cross-group connections based on Covariate-assisted Spectral Clustering. Panel A reports average return inferred connections across sample period. Panel B and Panel C report algorithm inferred connections and proof types inferred connections respectively. Connections are defined as

$$\textit{Within-Group Connection}_i = \frac{\#\text{ of Degrees of Coins within Group } i}{4N_i},$$
$$\textit{Cross-Group Connection}_i = \frac{\#\text{ of Degrees of Coins between Group } i \text{ and other Groups}}{4\bar{N}_i}.$$

*, **, and *** indicates statistical significance at the 10%, 5% and 1% levels respectively.

	Return			Algorithm			Proof Types		
	Within	Cross	Diff.	Within	Cross	Diff.	Within	Cross	Diff.
G1	0.137	0.110	0.027***	0.261	0.111	0.150	0.692	0.072	0.620
G2	0.144	0.113	0.031***	0.379	0.163	0.216	0.660	0.198	0.462
G3	0.046	0.065	-0.019***	0.807	0.151	0.656	0.622	0.046	0.576
G4	0.132	0.111	0.020***	0.071	0.129	-0.057	0.829	0.223	0.606
G5	0.107	0.103	0.004***	0.179	0.175	0.004	0.207	0.217	-0.010
All	0.113	0.101	0.013***	0.339	0.146	0.194	0.602	0.151	0.451

within- and cross-group centrality scores for both algorithm and proof types in Table 3 are all significantly positive comparing to Table 2. These results indicate that fundamental information introduces extra dimension of commonality for classify cryptocurrencies, and it improves information extraction from return dynamics by emphasizing the content behind the fundamental commonality induced return comovement.

Given the economic meanings of our grouping results, we now try to deepen the understanding our classification results by studying its asset pricing inference. We explore how to utilize our grouping information to make profit from a portfolio manager?s perspective. We initiate our tests from two angles with one based on the rational information diffusion channel and the other one based on behavioural bias interpretation. For testing information diffusion channel, we apply a similar testing procedure in the equities market (Hong et al., 2007; Rapach et al., 2015; Menzly and Ozbas, 2010) to the cryptocurrencies market. We have found limited evidence to support the information diffusion interpretation. Specifically, in the equities market, the cross-industry information is known to significantly predicts future returns of other industries, while it does not hold for the cryptocurrencies market. Actually, the cross-group information does not show any significant return predictability in cryptocurrencies market by Fama-MacBeth regression. This is not because the market is efficient enough to reflect all the information immediately, but it is a result of the fact that the market is crowded with sentiment that the fundamental information is far away from being priced. Although many investors apply the blockchain technology to their business, no one knows how to price these technologies, thus making it difficult to provide an explanation through information diffusion channel.

Therefore, we turn to the alternative channel, which focuses on invertors behavioural bias. As illustrated in Baker and Wurgler (2006), stocks that are newer, smaller, more volatile, less profitable, and those with analogous characteristics, hard to value or arbitrage, tend to suffer from strong sentiment bias or behavioural bias. Similarly, in the cryptocurrencies market, numerous literatures have documented sentiment effect (see Cretarola et al. (2017) for a comprehensive review.), and this motivates us to focus on the behaviour channel to study the asset pricing inference of our grouping results. We first examine the node covariate centrality score, which is defined as degree centrality of the

covariate matrix, and hypothesize that the fundamental centrality reflects the popularity of the fundamental settings of a group. Then, we argue that the investors who trade the coins in the group with a lower centrality score (less popular technologies) may face higher information asymmetry. The reason is that for the groups with special settings (the fundamental is less likely to be employed by other groups), investors have less peer fundamental information to assist understanding its price, which makes coins hard to value for investors. Besides, the liquidity for the altcoin market is much worse than the equities market so the arbitrage cost is very high. In this case, investors' speculation behaviour will create temporary price pressure which will result in a strong return reversal in the next trading day. Formally, we propose following hypothesis:

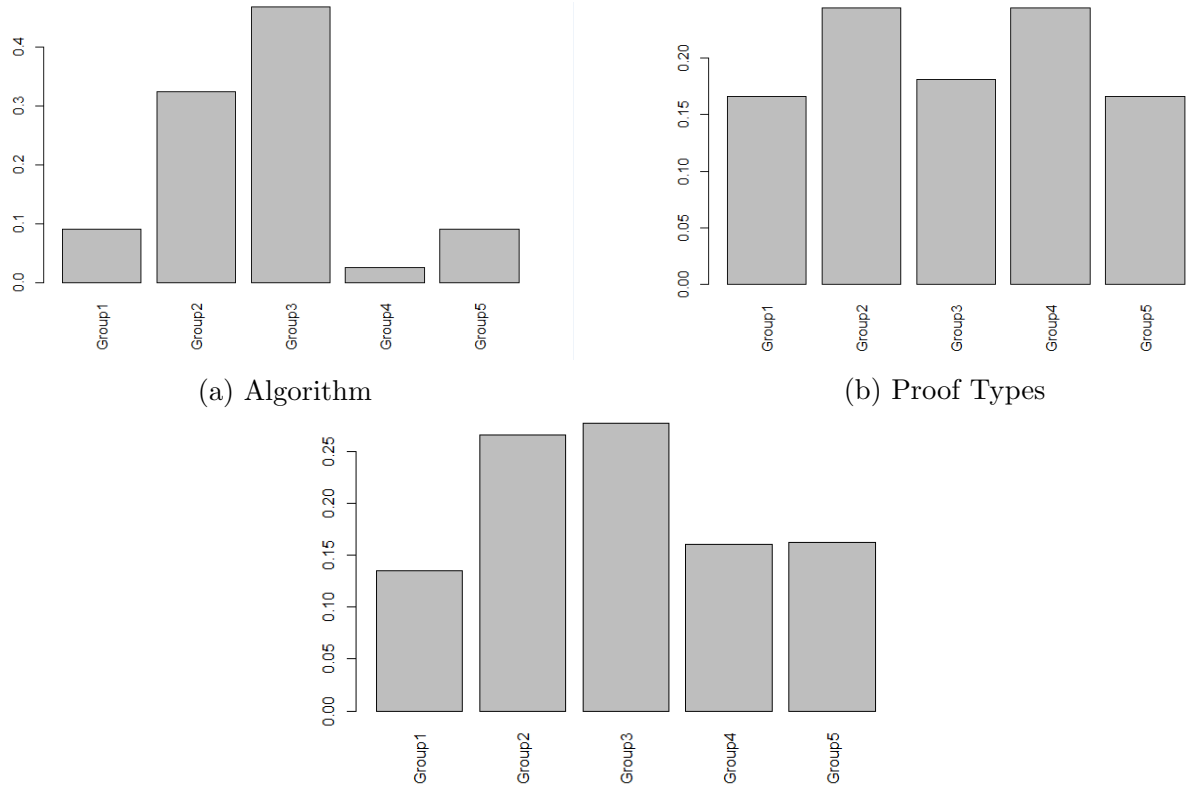
Hypothesis: *Contrarian strategy is more profitable for the groups with lower node covariate centrality scores.*

4.2 Group Centrality and Characteristic Distribution

From this section onwards, we will conduct several empirical tests to check the hypothesis proposed in the previous section. Firstly, as shown in Figure 5(c), Group 1 receives the lowest centrality score under a combined fundamental setting, and it indicates this group may have the most special settings comparing to other groups. Therefore, we may expect Group 1 to suffer from the most severe behavioural bias due to hard-to-value effect. By contrast, the centrality score of Group 3 is the highest, so it would has the most common fundamental settings, and thus the weakest return reversal is expected.

To verify the findings above, we then investigate the fundamental setting such as the algorithms and proof types among the five groups. Figure 6 plots the overall technology distribution of top 200 cryptocurrencies. Surprisingly, instead of SHA256 and Ethash (which are BTC and ETH fundamental algorithm respectively), "Multiple" algorithm[5] is the most widely used algorithm, (more than 35% of cryptocurrencies tend to use this new technology) in the current market according to Figure 6(a). In addition, Scrypt and X11

[5]Multiple is an algorithm that allows developers to mine any of the five used algorithms ? Scrypt, SHA256D, Qubit, Skein or Myriad-Groestl. Given this feature, it attracts more developers to contribute their computing power hence driving the developing of the market.



(a) Algorithm

(b) Proof Types

(c) Combined Fundamental

Figure 5: This figure depicts centrality score of each group in terms of fundamental settings. Subfigure (a) and (b) report centrality scores according to algorithm and proof type inferred network respectively, and (c) reports centrality score according to combined fundamental information. For the centrality score of combined fundamental information, we first construct the attribution matrix by aggregating both algorithm inferred adjacency matrix and proof type inferred adjacency matrix. We then calculate the degree centrality of each cryptocurrency and normalize the sum of centrality equals to 1. The group centrality is then defined as the average of centrality score of its group members.

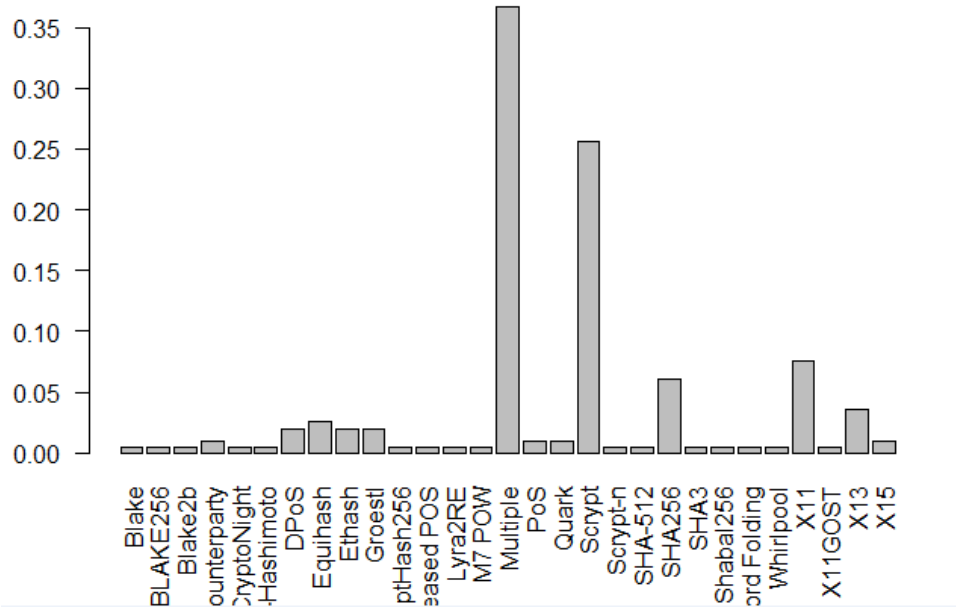
are the second and third most widely used algorithms. This suggests that new technologies developed in the cryptocurrencies market are more widely accepted by new start-ups. In terms of proof types, its distribution delivers a similar message. Although Proof-of-Work is still leading the other rewarding systems, mFBA rewarding system has become the second most widely used rewarding system, which is designed to fit Multiple algorithm, indicating that there is a trend of chasing new technologies in cryptocurrencies market.

In Figure 7, we present the evolution of technologies over the sampling period. Figure 7(a) plots the evolution of algorithm and 7(b) plots the proof types. Both figures indicate that the mainstream of fundamental technologies in cryptocurrencies market has become the Multiple algorithm plus mFBA proof type since 2017. Multiple algortihm is still below 18% in terms of total cryptocurrency support in the end of 2016 while it climbs to the first place with more than 36% of total supports in two years. Similarly, we find new rewarding mechanisms have been developed to fit the market demand, evidenced by the increase of mFBA market support. On the contrary, existing proof types, such as PoW and PoS lose their competitiveness in the recent days. So overall, from the figure, we witness that the market is growing fast with the support of new products and technologies.

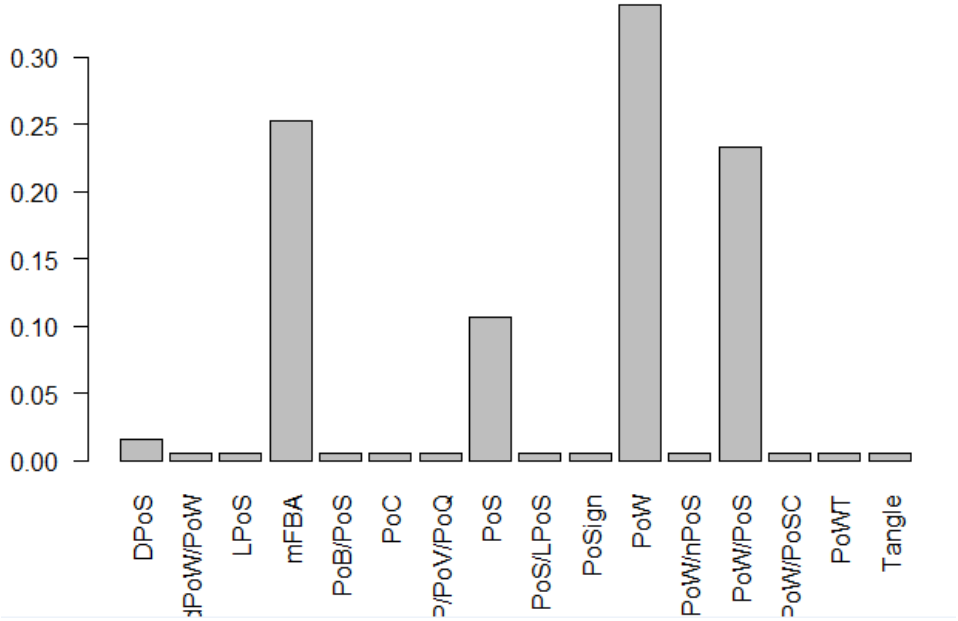
Figure 8 further illustrates the characteristic distribution of different groups. In particular, Group 2 and Group 3 concentrate their algorithms on Multiple and Scrypt respectively. Group 3, which has the highest fundamental centrality score, has more than 80% of group members employ Multiple algorithm, which is also the most popular technology in the market. Group 2 shows more than 60% of group members adopt Scrypt as their fundamental algorithm, which is the second most widely used techonology. Not surprisingly, these two groups achieve highest centrality score in terms of algorithm commonality in the market, evidenced by Figure 5(c). Interestingly, Group 1 members also adopts a common technology, Scrypt, as their main algorithm which explains why the algorithm centrality score of Group 1 is not the lowest. Neverthless, combining with proof types, Group 1 achieves the lowest fundamental centrality score. As shown in Figure 8(b), Group 1 uses mixed Proof-of-Work[6] and Proof-of-State[7] as their rewarding system while other groups

[6]PoW-based cryptocurrencies, such as bitcoin, uses mining, that is, the solving of computationally intensive puzzles to validate transactions and create new blocks.

[7]PoS-based cryptocurrencies, the creator of the next block is chosen via various combinations of random

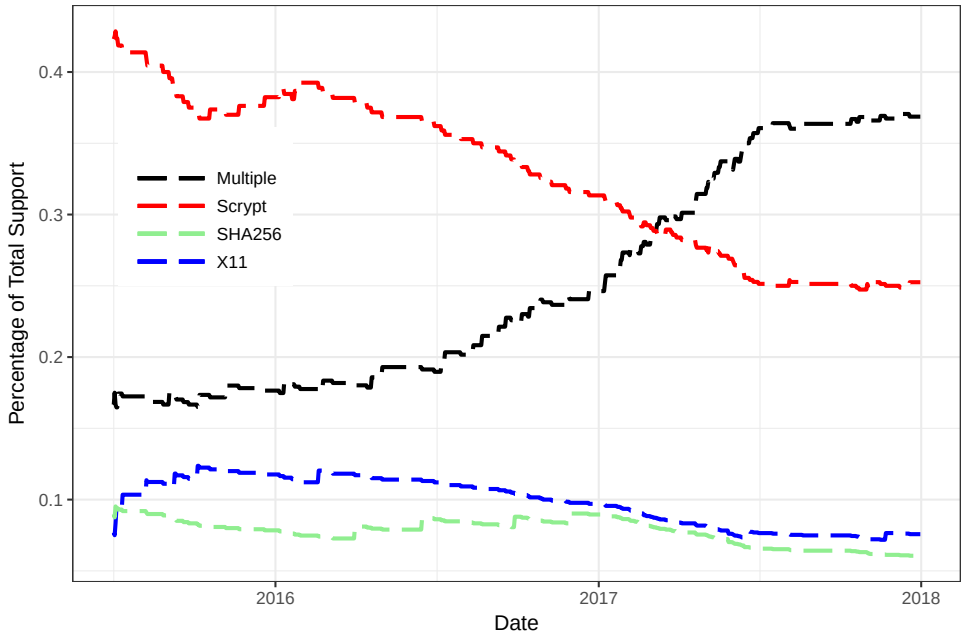


(a) Algorithm

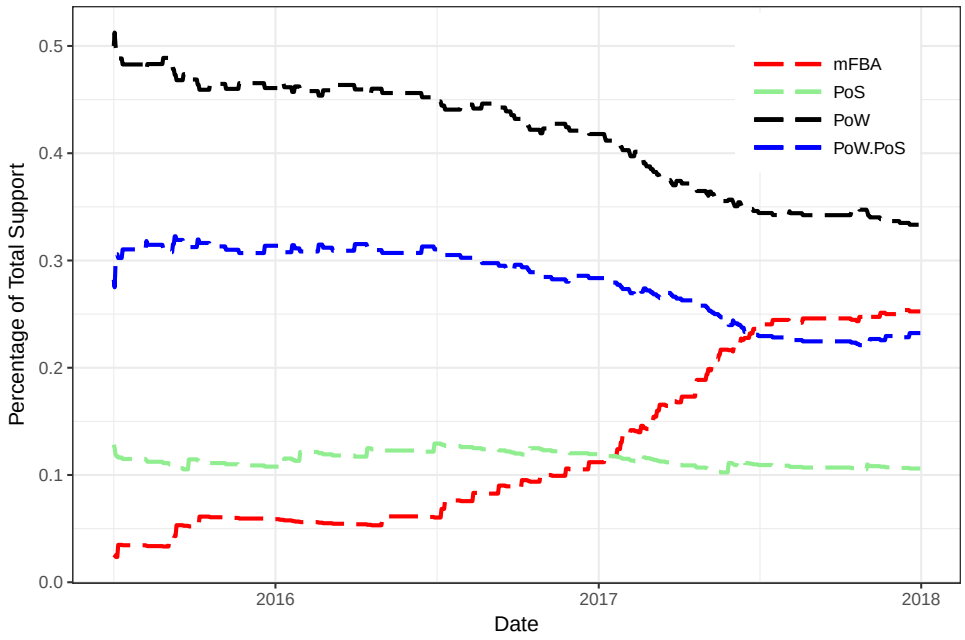


(b) Proof Types

Figure 6: This figure depicts the distribution of technology, including algorithm and proof types of cryptocurrencies market. In both figures, the y axis stands for the percentage of cryptocurrencies. While x axis stands for category of a fundamental technology. Subfigure (a) stands for the algorithm and (b) stands for the proof types.



(a) Algorithm



(b) Proof Types

Figure 7: This figure depicts the time variation of fundamental algorithms or proof types that are widely used in the cryptocurrencies market. In both figures, the y axis stands for the percentage of total cryptocurrencies. Subfigure (a) stands for the algorithm and (b) stands for the proof types.

mainly adopt single proof type as their main rewarding systems. Given PoW and PoS are two quite different rewarding systems, a mixed use in one Blockchain is preferred only in the early days while it becomes less common in nowadays as it may generate inconsistent objective functions among developers and hence discourage developers to do the mining.

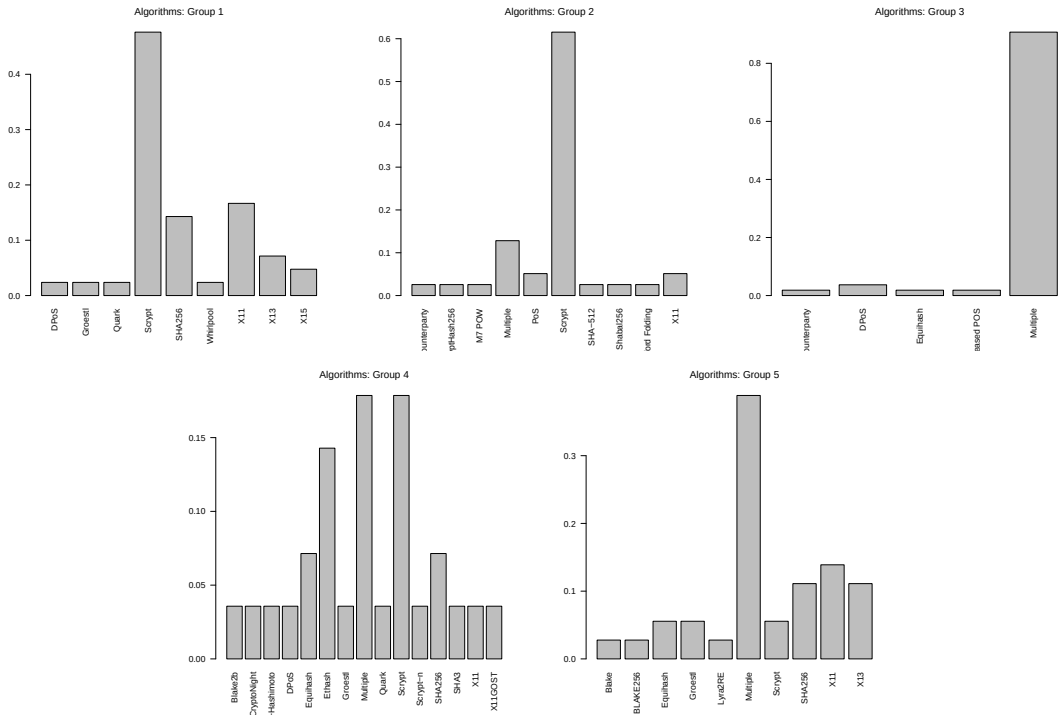
4.3 Cross-sectional Return Predictability

In this section, we test the hypothesis by checking the cross-sectional return predictability with a contrarian strategy. In equities market, contrarian strategy is well-designed to exploit return reversals by providing investors opportunities to achieve monthly abnormal returns of about 2% (Jegadeesh, 1990; Lehmann, 1990). We plot the cumulative returns of the whole cryptocurrencies market (all groups), the high centrality group (Group 3), the low centrality group (Group 1), and the median in Figure 9. Consistent with our hypothesis, Group 1 consistently enjoy the highest cumulative return based on the contrarian strategy while Group 3 receives the lowest cumulative return. Meanwhile, cumulative returns of contrarian strategy using all cryptocurrencies still exhibit a positive and upward trend located in between those of Group 1 and Group 3. In particular, the daily differences between the returns of low centrality group and whole market, whole market and high centrality group, and low centrality group and high centrality group, are as high as 1.25%, 7.67%, and 3.68%, respectively. In summary, the cross-sectional return predictability again provides strong evidence to support our hypothesis and reinforces the economic interpretation of our grouping results.

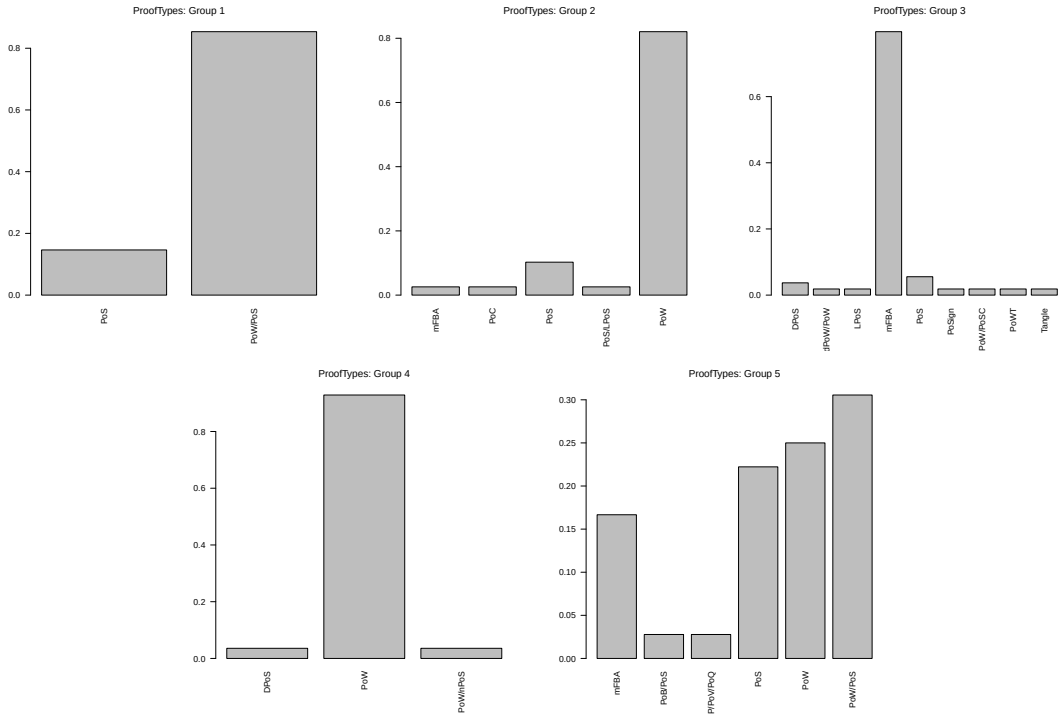
5 Conclusions

This paper studies the latent group structure in cryptocurrencies market and develops dynamic version of covariate-assisted spectral clustering methods to identify the group membership of each cryptocurrency. To obtain meaningful economic interpretations, we have proposed a hypothesis based on the investor behavioural bias channel, and we have the hypothesis tested by conducting asset pricing inference.

selection and wealth or age (i.e., the stake).



(a) Algorithm



(b) Proof Types

Figure 8: This figure depicts technology distribution of each group. In all figures, the y axis stands for the percentage of total group members. While x axis stands for category of a fundamental variable. Subfigure (a) is the algorithm and (b) is the proof types.

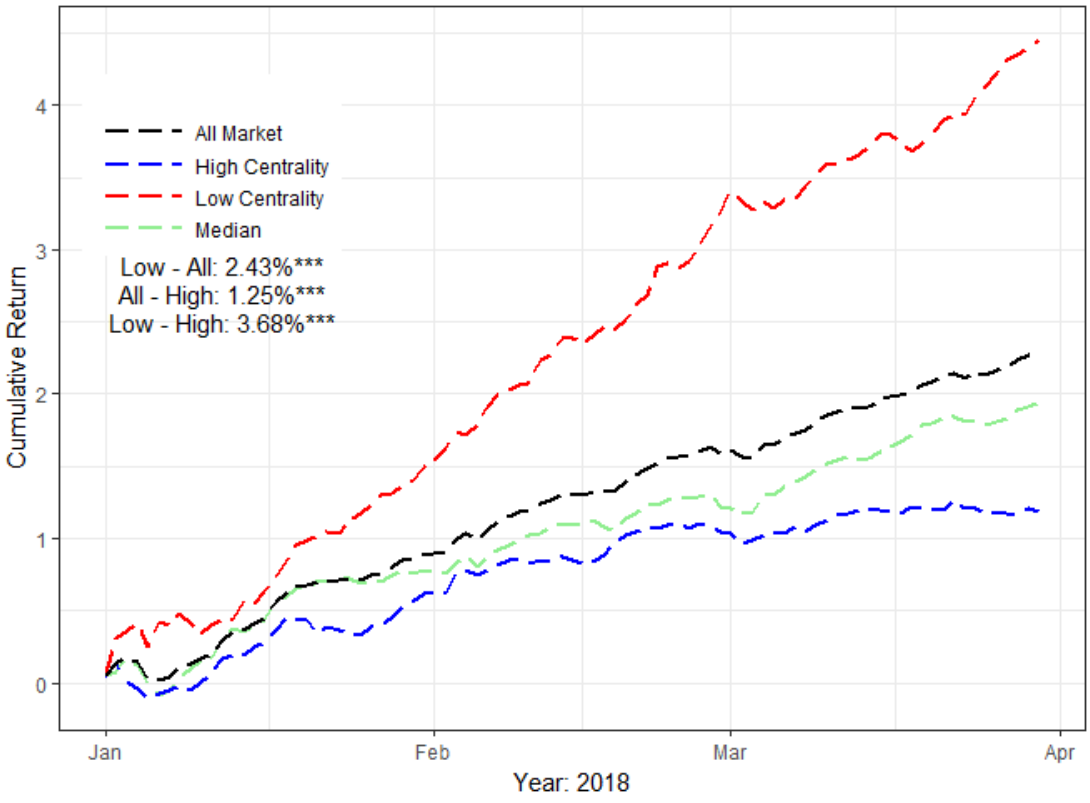


Figure 9: This figure plots cumulative returns of a contrarian strategy for high, median and low centrality groups. For each group, we conduct a equal weight daily contrarian strategy by shorting (longing) the group of cryptocurrencies with highest (lowest) return in previous trading day. We hold portfolio for 1 trading day and rebalance them at the close price of next trading day. We label group 3(1) as high (low) centrality group and the rest 3 groups as median centrality groups. The sample period is from 2018-01-01 to 2018-03-31.

Firstly, our classification results show that combining both return information and fundamental attributions of cryptocurrencies achieves a more consistent and economically meaningful classification results for analysis. The fundamental information indeed add on the return's information by providing more content for forming within-group connections.

Secondly, based on our clustering method, we show there is a "technology bias" in cryptocurrencies market, and we provide the explanation based on the fact that the investors face higher information uncertainty to trade cryptocurrencies with fewer peer fundamentals. Specifically, we propose the return reversal hypotheses through the investor behavioural bias channel. A contrarian strategy by shorting the cryptocurrencies with the highest return and longing the cryptocurrencies with lowest return in the previous trading day, achieves a 3.68% higher daily return in the lowest centrality group than that in the largest centrality group. This result complements the economic meanings of our grouping results and can be useful for investment applications.

APPENDICES

The notations that have been frequently used in the proofs are as follows: $[n] := \{1, 2, \cdots, n\}$ for any positive integer n, $\mathcal{M}_{m,n}$ be the set of all $m \times n$ matrices which have exactly one 1 and $n-1$ 0's in each row. $\mathbb{R}^{m\times n}$ denotes the set of all $m \times n$ real matrices. $\|\cdot\|$ is used to denote Euclidean ℓ_2-norm for vectors in $\mathbb{R}^{m\times 1}$ and the spectral norm for matrices on $\mathbb{R}^{m\times n}$. $\|\cdot\|_\infty$ denotes the largest element of the matrix in absolute value. $\|\cdot\|_F$ is the Frobenius norm on $\mathbb{R}^{m\times n}$, namely $\|M\|_F := \sqrt{\mathrm{tr}(M^\top M)}$. $\|\cdot\|_{\phi_2}$ is the sub-gaussian norm such that for any random variable x, there is $\|x\|_{\phi_2} := \sup_{\kappa\geq 1} \kappa^{-1/2}(\mathbb{E}|x|^\kappa)^{1/\kappa}$. $\mathbf{1}_{m,n} \in \mathbb{R}^{m\times n}$ consists of all 1's, ι_n denotes the column vector with n elements of all 1's. $\mathbb{1}_A$ denotes the indicator function of the event A.

A Preliminary Lemmas

Lemma 1. *Suppose A_t and X are the adjacency matrix and the node covariate matrices sampled from the SC-DCBM. Recall W_t and $\mathcal{W}_t$ are empirical and population weight matrices. Then, we have*

$$\sup_t \|W_t - \mathcal{W}_t\|_\infty = O_p(\xi),$$

where $\xi = \max(\sigma^2\|L_\tau\|_F\sqrt{\ln(TR)}, \sigma^2\|L_\tau\|\ln(TR), NRJ^2/\underline{\delta})$ and $\underline{\delta} = \inf_t\{\min_i \mathcal{D}_{\tau,t}(i,i)\}$.

Proof. Define $\mathcal{I}_t = \mathcal{X}L_{\tau,t}\mathcal{X}$. Then we have

$$\sup_t \|W_t - \mathcal{W}_t\|_\infty \leq \sup_t \|W_t - \mathcal{I}_t\|_\infty + \sup_t \|\mathcal{I}_t - \mathcal{W}_t\|_\infty.$$

For the first part, define $L_\tau = \sup_t L_{\tau,t}$ and $\zeta = \max(\sigma^2\|L_\tau\|_F\sqrt{\ln(TR)}, \sigma^2\|L_\tau\|\ln(TR))$, then by Hansen-Wright inequality (c.f., Theorem 1.1 of Rudelson and Vershynin (2013)), we have

$$\begin{aligned}
\Pr(\sup_t \|X^\top L_{\tau,t}X - \mathcal{X}^\top L_{\tau,t}\mathcal{X}\| > \zeta) &\leq \sum_{t=1}^{T} \Pr(\|X^\top L_\tau X - \mathcal{X}^\top L_\tau \mathcal{X}\| > \zeta) \\
&\leq 2T\exp\left\{-c\min\left(\frac{\zeta^2}{\sigma^4\|L_\tau\|_F^2}, \frac{\zeta}{\sigma^2\|L_\tau\|}\right)\right\} \\
&= O(1/R).
\end{aligned}$$

Next, denote $\mathcal{C}_t = \mathcal{D}_{\tau,t}^{-1/2} A_t \mathcal{D}_{\tau,t}^{-1/2}$, then we can decompose the second part into two parts:

$$\sup_t \|\mathcal{I}_t - \mathcal{W}_t\|_\infty = \sup_t \|\mathcal{X}(L_{\tau,t} - \mathcal{L}_{\tau,t})\mathcal{X}\|_\infty \leq \sup_t \|\mathcal{X}(L_{\tau,t} - \mathcal{C}_t)\mathcal{X}\|_\infty + \sup_t \|\mathcal{X}(\mathcal{C}_t - \mathcal{L}_{\tau,t})\mathcal{X}\|_\infty.$$

Then, for part one, we have

$$\begin{aligned}
\sup_t \|\mathcal{X}(L_{\tau,t} - \mathcal{C}_t)\mathcal{X}\|_\infty &= \sup_t \max_{s,r} \left| \sum_{i,j} \mathcal{X}_{is}\mathcal{X}_{jr} \frac{A_t(i,j)}{\sqrt{\mathcal{D}_{\tau,t}(i,i)\mathcal{D}_{\tau,t}(j,j)}} \left(\frac{\sqrt{\mathcal{D}_{\tau,t}(i,i)\mathcal{D}_{\tau,t}(j,j)}}{\sqrt{D_{\tau,t}(i,i)D_{\tau,t}(j,j)}} - 1 \right) \right| \\
&\leq \frac{1}{\delta} \max_{s,r} \sum_{i,j} |\mathcal{X}_{is}\mathcal{X}_{jr}| \sup_t \left\{ \max \left(\left| \frac{\mathcal{D}_{\tau,t}(i,i)}{D_{\tau,t}(i,i)} - 1 \right|, \left| \frac{\mathcal{D}_{\tau,t}(j,j)}{D_{\tau,t}(j,j)} - 1 \right| \right) \right\} \\
&= \max_{s,r} \sum_{i,j} |\mathcal{X}_{is}\mathcal{X}_{jr}| O_p(\underline{\delta}^{-3/2} \ln(TR)) \\
&= O_p\left(\frac{NRJ^2}{\underline{\delta}^{3/2}} \ln(TR) \right),
\end{aligned}$$

where the second to the last equality comes from the following proof. For any $i \in \{1, \cdots, N\}$ and $\varsigma = \underline{\delta}^{-1/2} \ln(TR)$, from Bernstein inequality,

$$\begin{aligned}
\Pr\left(\sup_t \left| \frac{D_{\tau,t}(i,i)}{\mathcal{D}_{\tau,t}(i,i)} - 1 \right| > \varsigma \right) &\leq \sum_{t=1}^{T} \Pr\left(\left| \frac{D_{\tau,t}(i,i)}{\mathcal{D}_{\tau,t}(i,i)} - 1 \right| > \varsigma \right) \\
&\leq 2T \exp\left\{ -\frac{\varsigma^2 \mathcal{D}_{\tau,t}(i,i)}{2 + \frac{2}{3}\varsigma} \right\} \\
&\leq 2T \exp\left\{ -\frac{\varsigma^2 \underline{\delta}}{2 + \frac{2}{3}\varsigma} \right\} \\
&= O(1/R).
\end{aligned}$$

For part two, similarly, we have

$$\begin{aligned}
\sup_t \|\mathcal{X}(\mathcal{C}_t - \mathcal{L}_{\tau,t})\mathcal{X}\|_\infty &= \sup_t \max_{s,r} \left| \sum_{i,j} \mathcal{X}_{is}\mathcal{X}_{jr} \frac{A_t(i,j) - \mathcal{A}_t(i,j)}{\sqrt{\mathcal{D}_{\tau,t}(i,i)\mathcal{D}_{\tau,t}(j,j)}} \right| \\
&\leq \max_{s,r} \left| \sum_{i,j} \mathcal{X}_{is}\mathcal{X}_{jr} \right| \sup_t \max_{i,j} \left| \frac{A_t(i,j) - \mathcal{A}_t(i,j)}{\sqrt{\mathcal{D}_{\tau,t}(i,i)\mathcal{D}_{\tau,t}(j,j)}} \right| \\
&= O_p\left(\frac{NRJ^2}{\underline{\delta}} \right).
\end{aligned}$$

Note that $\varsigma \to 0$ as $\underline{\delta}, R \to \infty$, we then know

$$\sup_t \|\mathcal{I}_t - \mathcal{W}_t\|_\infty = O_p\left(\frac{NRJ^2}{\underline{\delta}} \right).$$

Thus, by union bounds, we obtain

$$\sup_t \|W_t - \mathcal{W}_t\|_\infty = O_p\left(\zeta + \frac{NRJ^2}{\underline{\delta}}\right) = O_p(\xi).$$

□

Lemma 2. *Under Assumption 4, for any $\epsilon > 0$, we have*

$$\sup_t \|S_t - \mathcal{S}_t\| \leq (4 + c_w)\left\{\frac{3\ln(8NT/\epsilon)}{\underline{\delta}}\right\}^{1/2}, \tag{17}$$

with probability at least $1 - \epsilon$.

Proof. Note by triangular inequality, we have

$$\sup_t \|S_t - \mathcal{S}_t\| \leq \sup_t \left\|\alpha_t X W_t X^\top - \alpha_t \mathcal{X}\mathcal{W}_t\mathcal{X}^\top\right\| \tag{18}$$

$$+ \sup_t \left\|\mathcal{D}_{\tau,t}^{-1/2} A_t \mathcal{D}_{\tau,t}^{-1/2} - \mathcal{D}_{\tau,t}^{-1/2}\mathcal{A}_t\mathcal{D}_{\tau,t}^{-1/2}\right\| \tag{19}$$

$$+ \sup_t \left\|D_{\tau,t}^{-1/2} A_t D_{\tau,t}^{-1/2} - \mathcal{D}_{\tau,t}^{-1/2} A_t \mathcal{D}_{\tau,t}^{-1/2}\right\|. \tag{20}$$

For equation (18), we have,

$$\begin{aligned}
\sup_t \left\|\alpha_t X W_t X^\top - \alpha_t \mathcal{X}\mathcal{W}_t\mathcal{X}^\top\right\| &= \sup_t \left\|\alpha_t X(W_t - \mathcal{W}_t)X^\top\right\| + \sup_t \left\|\alpha_t X\mathcal{W}_t X^\top - \alpha_t\mathcal{X}\mathcal{W}_t\mathcal{X}^\top\right\| \\
&\leq \alpha_{\max} NRJ^2 \sup_t \|W_t - \mathcal{W}_t\| + 2\alpha_{\max} NRJ^2 \sup_t \|\mathcal{W}_t\| \\
&= O_p(\alpha_{\max} NRJ^2\xi).
\end{aligned}$$

So, by Assumption 4 we know, for large enough N, with probability at least $1 - \epsilon/2$,

$$\sup_t \left\|\alpha_t X W_t X^\top - \alpha_t \mathcal{X}\mathcal{W}_t\mathcal{X}^\top\right\| \leq c_w a$$

For equation (19), let $Y_t(i,j) = \mathcal{D}_{\tau,t}^{-1/2}[(A_t(i,j) - p_t(i,j))E_{ij}]\mathcal{D}_{\tau,t}^{-1/2}$ with $E_{ij} \in \mathbb{R}^{N\times N}$ being the matrix with 1 in ij and ji'th positions and 0 everywhere else. Then we know

$$\sup_t \|Y_t(i,j)\| \leq \sup_t \sqrt{\mathcal{D}_{\tau,t}(i,i)\mathcal{D}_{\tau,t}(j,j)} \leq \frac{1}{\underline{\delta}}, \quad v^2 = \sup_t \|\sum \mathbb{E}(Y_t^2(i,j))\| \leq \frac{1}{\underline{\delta}}.$$

So, denote $a = \left\{\frac{3\ln(8NT/\epsilon)}{\underline{\delta}}\right\}^{1/2}$, which is smaller than 1 by assumption, and by matrix

Bernstein inequality, we have

$$\begin{aligned}\Pr(\sup_t \|\mathcal{D}_{\tau,t}^{-1/2}[A_t(i,j)-\mathcal{A}_t(i,j)]\mathcal{D}_{\tau,t}^{-1/2}\| > a) &\le \sum_{t=1}^{T} \Pr(\|\mathcal{D}_{\tau,t}^{-1/2}[A_t(i,j)-\mathcal{A}_t(i,j)]\mathcal{D}_{\tau,t}^{-1/2}\| > a) \\ &\le 2NT\exp\left(-\frac{a^2}{2/\underline{\delta}+2a/3\underline{\delta}}\right) \\ &\le 2NT\exp\left(-\frac{3\ln(8NT/\epsilon)}{3}\right) \\ &= \epsilon/4.\end{aligned}$$

Hence, with probability at least $1-\epsilon/4$,

$$\sup_t \|\mathcal{D}_{\tau,t}^{-1/2} A_t \mathcal{D}_{\tau,t}^{-1/2} - \mathcal{D}_{\tau,t}^{-1/2}\mathcal{A}_t\mathcal{D}_{\tau,t}^{-1/2}\| \le a \tag{21}$$

Lastly, for equation (20), by Qin and Rohe (2013) and setting $\lambda = a\mathcal{D}_{\tau,t}(i,i)$ we have

$$\begin{aligned}\Pr(|D_{\tau,t}(i,i)-\mathcal{D}_{\tau,t}(i,i)| \ge \lambda) &\le \left(\exp\left\{-\frac{\lambda^2}{2\mathcal{D}_{\tau,t}(i,i)}\right\} + \exp\left\{-\frac{\lambda^2}{2\mathcal{D}_{\tau,t}(i,i)+\frac{2}{3}\lambda}\right\}\right) \\ &\le 2\exp\left\{-\frac{\lambda^2}{2\mathcal{D}_{\tau,t}(i,i)+\frac{2}{3}\lambda}\right\} \\ &= 2\exp\left\{-\frac{a^2\mathcal{D}_{\tau,t}(i,i)}{2+\frac{2}{3}a}\right\} \\ &\le 2\exp\left\{-\ln(8NT/\epsilon)\times\frac{\mathcal{D}_{\tau,t}(i,i)}{\underline{\delta}}\right\} \\ &\le \frac{\epsilon}{4NT}.\end{aligned}$$

Further note that

$$\begin{aligned}\Pr\left(\sup_t \|\mathcal{D}_{\tau,t}^{-1/2} D_{\tau,t}^{1/2} - I\| \ge a\right) &\le \sum_{t=1}^{T} \Pr\left(\|\mathcal{D}_{\tau,t}^{-1/2} D_{\tau,t}^{1/2} - I\| \ge a\right) \\ &\le \sum_{t=1}^{T} \Pr\left(\max_i \left|\frac{D_{\tau,t}(i,i)}{\mathcal{D}_{\tau,t}(i,i)} - 1\right| \ge a\right) \\ &\le \sum_{t=1}^{T}\sum_{i=1}^{N} \Pr\left(|D_{\tau,t}(i,i)-\mathcal{D}_{\tau,t}(i,i)| \ge a\mathcal{D}_{\tau,t}(i,i)\right) \\ &\le NT \times \frac{\epsilon}{4NT} \\ &= \epsilon/4.\end{aligned}$$

Therefore, with probability at least $1-\epsilon/4$, we have

$$\begin{aligned}\sup_t \|D_{\tau,t}^{-1/2}A_tD_{\tau,t}^{-1/2} - \mathcal{D}_{\tau,t}^{-1/2}A_t\mathcal{D}_{\tau,t}^{-1/2}\| &= \sup_t \|L_{\tau,t} - \mathcal{D}_{\tau,t}^{-1/2}D_{\tau,t}^{1/2}L_{\tau,t}D_{\tau,t}^{1/2}\mathcal{D}_{\tau,t}^{-1/2}\| \\ &= \sup_t \|(I-\mathcal{D}_{\tau,t}^{-1/2}D_{\tau,t}^{1/2})L_{\tau,t}D_{\tau,t}^{1/2}\mathcal{D}_{\tau,t}^{-1/2} + L_{\tau,t}(I-D_{\tau,t}^{1/2}\mathcal{D}_{\tau,t}^{-1/2})\| \\ &\leq \sup_t \|\mathcal{D}_{\tau,t}^{-1/2}D_{\tau,t}^{1/2} - I\| \sup_t \|\mathcal{D}_{\tau,t}^{-1/2}D_{\tau,t}^{1/2}\| + \sup_t \|\mathcal{D}_{\tau,t}^{-1/2}D_{\tau,t}^{1/2} - I\| \\ &\leq a^2 + 2a\end{aligned}$$

where the second last inequality comes from the fact that $\sup_t \|L_{\tau,t}\| \leq 1$.

Therefore, joining the results for these three equations, we have, with probability at least $1-\epsilon$,

$$\|S_t - \mathcal{S}_t\| \leq a^2 + 3a + c_w a \leq (4+c_w)a = (4+c_w)\left\{\frac{3\ln(8NT/\epsilon)}{\underline{\delta}}\right\}^{1/2}. \tag{22}$$

□

Lemma 3. *Under the dynamic SC-DCBM with K blocks, define $\Gamma_{\tau,t} \in \mathbb{R}^{N\times K}$ with columns containing the top K eigenvectors of $\mathcal{S}_t$. Then, under Assumption 4, there exists an orthogonal matrix U_t depending on τ_t for each $t = 1, \cdots, T$, such that for any $i, j = 1, \cdots, N$,*

$$\Gamma_{\tau,t} = \Psi_{\tau,t}^{1/2}Z_t(Z_t^\top\Psi_{\tau,t}Z_t)^{-1/2}U_t \;\; \textit{and} \;\; \Gamma_{\tau,t}^*(i,*) = \Gamma_{\tau,t}^*(j,*) \Longleftrightarrow Z_t(i,*) = Z_t(j,*),$$

where $\Gamma_{\tau,t}^*(i,*) = \Gamma_{\tau,t}(i,*)/\|\Gamma_{\tau,t}(i,*)\|$.

Proof. Denote $D_{B,t}$ as a diagonal matrix with entries $D_{B,t}(i,i) = \sum_{j=1}^K B_t(i,j)$, and $\Psi_{\tau,t} = \mathrm{Diag}(\psi_{\tau,t})$ with $\psi_{\tau,t}(i) = \psi_t\frac{\mathcal{D}_t(i,i)}{\mathcal{D}_{\tau,t}(i,i)}$. Then, Under the dynamic SC-DCBM, we have the decomposition below

$$\mathcal{L}_{\tau,t} = \mathcal{D}_{\tau,t}^{-1/2}\mathcal{A}_t\mathcal{D}_{\tau,t}^{-1/2} = \Psi_{\tau,t}^{1/2}Z_tB_{L,t}Z_t^\top\Psi_{\tau,t}^{1/2},$$

where $B_{L,t} = D_{B,t}^{-1/2}B_tD_{B,t}^{-1/2}$.

Define M_t such that $\mathcal{X} = \mathbb{E}(X) = \Psi_{\tau,t}^{1/2}Z_tM_t$, and $\Omega_t = B_{L,t} + \alpha_tM_t\mathcal{W}_tM_t^\top$, then we know

$$\mathcal{S}_t = \Psi_{\tau,t}^{1/2}Z_t\Omega_tZ_t^\top\Psi_{\tau,t}^{1/2}. \tag{23}$$

Now, denote $Y_{\tau,t} = Z_t^\top \Psi_{\tau,t} Z_t$, and let $H_{\tau,t} = Y_{\tau,t}^{1/2} \Omega_t Y_{\tau,t}^{1/2}$. Then, by eigen-decomposition, we have $H_{\tau,t} = U_t \Lambda_t U_t^\top$. Define $\Gamma_{\tau,t} = \Psi_{\tau,t}^{1/2} Z_t Y_{\tau,t}^{-1/2} U_t$, then

$$\begin{aligned}
\Gamma_{\tau,t}^\top \Gamma_{\tau,t} &= U_t^\top Y_{\tau,t}^{-1/2} Z_t^\top \Psi_{\tau,t}^{1/2} \Psi_{\tau,t}^{1/2} Z_t Y_{\tau,t}^{-1/2} U_t \\
&= U_t^\top Y_{\tau,t}^{-1/2} Y_{\tau,t} Y_{\tau,t}^{-1/2} U_t \\
&= U_t^\top U_t = I,
\end{aligned}$$

and we have

$$\begin{aligned}
\mathcal{S}_t \Gamma_{\tau,t} &= (\Psi_{\tau,t}^{1/2} Z_t \Omega_t Z_t^\top \Psi_{\tau,t}^{1/2}) \Psi_{\tau,t}^{1/2} Z_t (Z_t^\top \Psi_{\tau,t} Z_t)^{-1/2} U_t \\
&= \Psi_{\tau,t}^{1/2} Z_t \Omega_t Y_{\tau,t}^{1/2} U_t \\
&= \left\{ \Psi_{\tau,t}^{1/2} Z_t Y_{\tau,t}^{-1/2} \left(Y_{\tau,t}^{1/2} \Omega_t Y_{\tau,t}^{1/2} \right) \right\} U_t \\
&= \Psi_{\tau,t}^{1/2} Z_t Y_{\tau,t}^{-1/2} (U_t \Lambda_t U_t^\top) U_t \\
&= \Gamma_{\tau,t} \Lambda_t.
\end{aligned}$$

Following Qin and Rohe (2013), it is obvious that

$$\Gamma_{\tau,t}^*(i,*) = \frac{\Gamma_{\tau,t}(i,*)}{\|\Gamma_{\tau,t}(i,*)\|} = Z_{i,t} U_t.$$

Then, by directly applying the Lemma 1 in Binkiewicz et al. (2017), we complete the proof. □

B Main Proof

Proof. By Tao (2018), we can easily extends its result to

$$\sup_t \frac{|\mathbb{M}_t|}{N} \leq \frac{c_1(\varepsilon) K}{m_z^2 N \lambda_{K,\max}^2} \sup_t \left\| \widehat{\mathcal{S}}_{t,r} - \mathcal{S}_t \right\|^2. \tag{24}$$

Then, for $\mathcal{S}_t$, we have the following representation:

$$\mathcal{S}_t = \mathcal{D}_{\tau,t}^{-1/2} \Psi Z_t B_t Z_t^\top \Psi \mathcal{D}_{\tau,t}^{-1/2} + \alpha_t \mathcal{X} \mathcal{W}_t \mathcal{X}^\top, \tag{25}$$

To figure out the upper bound of the estimation error, we have to evaluate the error bound $\sup_t \left\| \widehat{\mathcal{S}}_{t,r} - \mathcal{S}_t \right\|$. Define

$$\mathcal{S}_{t,r} = \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} W_{r,l}(i) \mathcal{S}_{t+i}, \tag{26}$$

then by triangle inequality, we have

$$\Delta(r) = \sup_t \left\| \widehat{\mathcal{S}}_{t,r} - \mathcal{S}_t \right\| \leq \sup_t \left\| \widehat{\mathcal{S}}_{t,r} - \mathcal{S}_{t,r} \right\| + \sup_t \left\| \mathcal{S}_{t,r} - \mathcal{S}_t \right\| = \Delta_1(r) + \Delta_2(r). \tag{27}$$

For $\Delta_1(r)$, by Lemma 2, we have

$$\begin{aligned}
\Delta_1(r) &= \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} W_{r,l}(i) \sup_t \left\| S_{t+i} - \mathcal{S}_{t+i} \right\| \\
&\leq \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} W_{r,l}(i) \left\{ (4 + c_w) \left[\frac{3 \ln(8NT/\epsilon)}{\underline{\delta}} \right]^{1/2} \right\} \\
&\leq W_{\max}(4 + c_w) \left\{ \frac{3 \ln(8NT/\epsilon)}{\underline{\delta}} \right\}^{1/2}.
\end{aligned} \tag{28}$$

For Δ_2, we have the following decomposition

$$\Delta_2(r) = \sup_t \left\| \mathcal{S}_{t,r} - \mathcal{S}_t \right\| \leq \sup_t \left\| \mathcal{S}_{t,r} - \widetilde{\mathcal{S}}_{t,r} \right\| + \sup_t \left\| \widetilde{\mathcal{S}}_{t,r} - \mathcal{S}_t \right\| = \Delta_{21} + \Delta_{22}, \tag{29}$$

where

$$\widetilde{\mathcal{S}}_{t,r} = \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} W_{r,l}(i) \left(\mathcal{D}_{\tau,t}^{-1/2} Z_t B_{t+i} Z_t^\top \mathcal{D}_{\tau,t}^{-1/2} + \alpha_{t+i} \mathcal{X} \mathcal{W}_t \mathcal{X}^\top \right). \tag{30}$$

Then for Δ_{21}, we have

$$\begin{aligned}
\Delta_{21,t} &\leq W_{\max} \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} \sup_t \left\| \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_{t+i} B_{t+i} Z_{t+i}^\top \Psi \mathcal{D}_{\tau,t+i}^{-1/2} - \mathcal{D}_{\tau,t}^{-1/2} \Psi Z_t B_{t+i} Z_t^\top \Psi \mathcal{D}_{\tau,t}^{-1/2} \right\| \\
&\leq W_{\max} \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} \sup_t \left\{ \left(\left\| \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_{t+i} \right\| + \left\| \mathcal{D}_{\tau,t}^{-1/2} \Psi Z_t \right\| \right) \|B_{t+i}\| \left\| \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_{t+i} - \mathcal{D}_{\tau,t}^{-1/2} \Psi Z_t \right\| \right\} \\
&\leq W_{\max} \frac{1}{|\mathcal{F}_r|} \sum_{i \in \mathcal{F}_r} \sup_t \left\{ \left(\left\| \mathcal{D}_{\tau,t}^{-1/2} \right\| \|Z_{t+i}\| + \left\| \mathcal{D}_{\tau,t}^{-1/2} \right\| \|Z_t\| \right) \|B_{t+i}\| \left\| \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_{t+i} - \mathcal{D}_{\tau,t}^{-1/2} \Psi Z_t \right\| \right\},
\end{aligned}$$

where the last inequality comes from the fact that $\|\Psi\| = \max_i \left| \sqrt{\psi_i} \right| \leq 1$.

Then, observe that $\sup_t \left\| \mathcal{D}_{\tau,t}^{-1/2} \right\| \leq \underline{\delta}^{-1/2}$, $\sup_t \|Z_t\| \leq P_{\max}^{1/2}$, $\sup_t \|B_t\| \leq K$, we then have

$$\sup_t \left\{ \left\| \mathcal{D}_{\tau,t}^{-1/2} \right\| \|Z_{t+i}\| + \left\| \mathcal{D}_{\tau,t}^{-1/2} \right\| \|Z_t\| \right\} \leq 2 \underline{\delta}^{-1/2} P_{\max}^{1/2}. \tag{31}$$

Further, note that

$$\begin{aligned}
\sup_t \left\| \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_{t+i} - \mathcal{D}_{\tau,t}^{-1/2} \Psi Z_t \right\| &\leq \sup_t \left\{ \left\| \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_{t+i} - \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_t \right\| + \left\| \mathcal{D}_{\tau,t+i}^{-1/2} \Psi Z_t - \mathcal{D}_{\tau,t}^{-1/2} \Psi Z_t \right\| \right\} \\
&\leq \sup_t \left\{ \left\| \mathcal{D}_{\tau,t+i}^{-1/2} \right\| \|\Psi\| \|Z_{t+i} - Z_t\| + \left(\left\| \mathcal{D}_{\tau,t+i}^{-1/2} \right\| \|\Psi\| + \left\| \mathcal{D}_{\tau,t}^{-1/2} \right\| \|\Psi\| \right) \|Z_t\| \right\} \\
&\leq \sqrt{\frac{2|r|s}{\underline{\delta}}} + \sqrt{\frac{4P_{\max}}{\underline{\delta}}}.
\end{aligned}$$

Then, combine the results above with the assumption $\underline{\delta} > 3\ln(8NT/\epsilon)$ in Lemma 2, we have

$$\Delta_{21} \leq \frac{2W_{\max}K}{\sqrt{3\ln(8NT/\epsilon)}}(\sqrt{2P_{\max}rs} + 2P_{\max}). \tag{32}$$

Lastly, for Δ_{22}, similarly, we define

$$\widetilde{\mathcal{S}}_{t,r} = \frac{1}{|\mathcal{F}_r|}\sum_{i\in\mathcal{F}_r} W_{r,l}(i)\left(Y_{\tau,t}B_{t+i}Y_{\tau,t}^\top + \alpha_{t+i}\mathcal{X}\mathcal{W}_t\mathcal{X}^\top\right). \tag{33}$$

and $Y_{\tau,t} := \mathcal{D}_{\tau,t}^{-1/2}\Psi Z_t$.

Then, apply the results in Pensky and Zhang (2017) and proof of Lemma 2, we obtain

$$\begin{aligned}
\Delta_{22} &= \sup_t \left\|\widetilde{\mathcal{S}}_{t,r} - \mathcal{S}_t\right\| \\
&= \frac{1}{|\mathcal{F}_r|}\sum_{i\in\mathcal{F}_r} W_{r,l}(i)\sup_t\left(Y_{\tau,t}\left\|B_{t+i}-B_t\right\|Y_{\tau,t}^\top + \left\|\alpha_{t+i}-\alpha_t\right\|\left\|\mathcal{X}\mathcal{W}_t\mathcal{X}^\top\right\|\right) \\
&\leq \sup_t\left\{\max_{1\leq j'\leq N}\sum_{j=1}^{N}\left|(Y_{\tau,t}Q_{r,t}Y_{\tau,t}^\top)(j,j')\right|\right\} + 2\alpha_{\max}W_{\max}NRJ^2\sup_t\|\mathcal{W}_t\| \\
&\leq \sup_t\left\{\max_{k,k'}|Q_{r,t}|\max_{1\leq j'\leq N}\sum_{k=1}^{K}\sum_{k'=1}^{K}\left[\sum_{j\in\mathcal{G}_{t,k}}Y_{\tau,t}(j,k)\right]Y_{\tau,t}(j',k')\right\} + c_w\left\{\frac{3\ln(8NT/\epsilon)}{\underline{\delta}}\right\}^{1/2} \\
&\leq \frac{NLW_{\max}}{\underline{\delta}\cdot l!}\left(\frac{r}{T}\right)^\beta + c_w\left\{\frac{3\ln(8NT/\epsilon)}{\underline{\delta}}\right\}^{1/2} \\
&\leq \frac{NLW_{\max}}{3\ln(8NT/\epsilon)\cdot l!}\left(\frac{r}{T}\right)^\beta + c_w\left\{\frac{3\ln(8NT/\epsilon)}{\underline{\delta}}\right\}^{1/2}
\end{aligned} \tag{34}$$

where the last two inequalities come from the fact that $\max_i\psi_i \leq 1$ and $\underline{\delta} > b^2$.

Now, combine the results provided by equation (24), (28), (32), and (34), we derive the upper bound for misclustering rate of dynamic DCBM: with probability at least $1-\epsilon$,

$$\sup_t\frac{|\mathbb{M}_t|}{N} \leq \frac{c(\varepsilon)KW_{\max}^2}{m_z^2N\lambda_{K,\max}^2}\left\{(4+2c_w)\frac{b}{\underline{\delta}^{1/2}} + \frac{2K}{b}(\sqrt{2P_{\max}rs}+2P_{\max}) + \frac{NL}{b^2\cdot l!}\left(\frac{r}{T}\right)^\beta\right\}^2.$$

where $b = \sqrt{3\ln(8NT/\epsilon)}$, $\lambda_{K,\max} = \max_t\{\lambda_{K,t}\}$ and $c(\varepsilon) = 2^9(2+\varepsilon)^2$. □

References

Baker, M. and J. Wurgler (2006): "Investor sentiment and the cross-section of stock returns," *The Journal of Finance*, 61, 1645–1680.

Berkman, H., P. D. Koch, L. Tuttle, and Y. J. Zhang (2012): "Paying attention: overnight returns and the hidden cost of buying at the open," *Journal of Financial and Quantitative Analysis*, 47, 715–741.

Bhattacharyya, S. and S. Chatterjee (2017): "Spectral clustering for dynamic stochastic block model," *Working Paper*.

Binkiewicz, N., J. T. Vogelstein, and K. Rohe (2017): "Covariate-assisted spectral clustering," *Biometrika*, 104, 361–377.

Black, F. (1986): "Noise," *Journal of Financial Service*, 41, 529–543.

Campbell, JY., a. G. S. and J. Wang (1993): "Trading volume and serial correlation in stock returns," *Quarterly Journal of Economics*, 108, 905–939.

Chaudhuri, K., F. Chung, and A. Tsiatas (2012): "Spectral clustering of graphs with general degrees in the extended planted partition model," in *Conference on Learning Theory*, 35–1.

Chen, C., W. Härdle, H. AJ, and W. W (2018): "Pricing Crypto Currency Options, the case of CRIX," *Journal of Financial Econometrics*.

Chen, K. and J. Lei (2017): "Network cross-validation for determining the number of communities in network data," *Journal of the American Statistical Association*, 1–11.

Cretarola, A., G. Figá-Talamanca, and M. Patacca (2017): "A Sentiment-Based Model for the Bitcoin: Theory, Estimation and Option Pricing," *Working Paper*.

Detzel, A. L., H. Liu, J. Strauss, G. Zhou, and Y. Zhu (2018): "Bitcoin: Learning, Predictability and Profitability via Technical Analysis," .

Elendner, H., S. Trimborn, B. Ong, D. Lee, and T.M (2015): *The Cross-Section of Crypto-currencies as Financial Assets: Investing in Crypto-currencies beyond Bitcoin*, Elsevier, vol. 1, 145–173.

Grossman, S. and M. Miller (1988): "Liquidity and market structure," *Journal of Finance*, 43, 617–633.

Hoberg, G. and G. Phillips (2016): "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, 124, 1423–1465.

Hong, H., W. Torous, and R. Valkanov (2007): "Do industries lead stock markets?" *Journal of Financial Economics*, 83, 367–396.

Jegadeesh, N. (1990): "Evidence of predictable behavior of security returns," *The Journal of finance*, 45, 881–898.

Jegadeesh, N. and S. Titman (1995): "Short-horizon return reversals and the bid-ask spread," *Journal of Financial Intermediation*, 4, 116–132.

Karrer, B. and M. E. J. Newman (2011): "Stochastic blockmodels and community structure in networks," *Physical Review E*, 83, 016107.

Lee, D., K. Chuen, G. Li, and W. Yu (2017): "Cryptocurrency: A New Investment Opportunity?" *Working Paper*.

Lehmann, B. N. (1990): "Fads, martingales, and market efficiency," *The Quarterly Journal of Economics*, 105, 1–28.

Lei, J. and A. Rinaldo (2015): "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, 43, 215–237.

Li, T., E. Levina, and J. Zhu (2016): "Network cross-validation by edge sampling," *arXiv preprint arXiv:1612.04717*.

Menzly, L. and O. Ozbas (2010): "Market segmentation and cross-predictability of returns," *The Journal of Finance*, 65, 1555–1580.

Nakamoto, S. (2008): "Bitcoin: A peer-to-peer electronic cash system," .

Ong, B., T. M., L. Guo, and K. David Lee (2015): *Evaluating the Potential of Alternative Cryptocurrencies*, Elsevier, Academic Press, 81–135.

Pensky, M. and T. Zhang (2017): "Spectral clustering in the dynamic stochastic block model," *arXiv preprint arXiv:1705.01204*.

Qin, T. and K. Rohe (2013): "Regularized spectral clustering under the degree-corrected stochastic blockmodel," in *Advances in Neural Information Processing Systems*, 3120–3128.

Rapach, D., J. Strauss, J. Tu, and G. Zhou (2015): "Industry interdependencies and cross-industry return predictability," *Working Paper*.

Rudelson, M. and R. Vershynin (2013): "Hanson-wright Inequality and Sub-gaussian Concentration," *Electronic Communications in Probability*, 18.

Shiller, R. (1984): "Stock prices and social dynamics activity," *Brookings Papers Econom*, 12, 457–498.

Subrahmanyam, A. (2005): "Distinguishing between rationales for short-horizon predictability of stock returnsg," *Financial Review*, 40, 11–35.

Tao, Y. (2018): "Covariate-assisted Spectral Clustering in Dynamic Networks," *arXiv preprint arXiv:1802.03708*.

Trimborn, S.and Li, M. and W. Härdle (2017a): "A liquidity constrained investment approach," *revise and resubmit Journal of Financial Econometrics*.

Trimborn, S. and W. Härdle (2017b): "CRIX an index for blockchain based currencies," *Working Paper*.

Wang, Y. X. R. and P. J. Bickel (2017): "Likelihood-based model selection for stochastic block models," *The Annals of Statistics*, 45, 500–528.

Zhang, Y., M. Poux-Berthe, C. Wells, K. Koc-Michalska, and K. Rohe (2017): "Discovering Political Topics in Facebook Discussion threads with Spectral Contextualization," *arXiv preprint arXiv:1708.06872*.

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101, 1418–1429.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.

021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.

022 " Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.

023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.

024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.

025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.

026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.

027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.

028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.

030 "Gaussian Process Forecast with multidimensional distributional entries" by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018.

031 "Instrumental variables regression" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

032 "Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective" by Li Guo, Yubo Tao and Wolfgang Karl Härdle, July 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.