Received: 14 October 2015 | Revised: 5 April 2018

DOI: 10.1002/jae.2642

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Realized networks

Christian Brownlees[1] | Eulàlia Nualart[1] | Yucheng Sun[2]

[1]Department of Economics and Business, Universitat Pompeu Fabra and Barcelona GSE, Barcelona, Spain

[2]International School of Economics and Management, Capital University of Economics and Business, Beijing, China

Correspondence
Yucheng Sun, International School of Economics and Management, Capital University of Economics and Business, 121 Zhangjialukou, Huaxiang Fengtai District, Beijing 100070, China.
Email: cueb_sun@163.com

Funding information
Spanish Ministry of Science and Technology, Grant/Award Number: MTM2012-37195; Analysis of Big Data in Economics; Empirical Applications 2016 BBVA Foundation Grants for Scientific Research Teams; European Union FP7-PEOPLE-2012-CIG, Grant/Award Number: 333938

Summary

We introduce LASSO-type regularization for large-dimensional realized covariance estimators of log-prices. The procedure consists of shrinking the off-diagonal entries of the inverse realized covariance matrix towards zero. This technique produces covariance estimators that are positive definite and with a sparse inverse. We name the estimator realized network, since estimating a sparse inverse realized covariance matrix is equivalent to detecting the partial correlation network structure of the daily log-prices. The large sample consistency and selection properties of the estimator are established. An application to a panel of US blue chip stocks shows the advantages of the estimator for out-of-sample GMV asset allocation.

1 | INTRODUCTION

The covariance matrix of the log-prices of financial assets is a fundamental ingredient in many applications ranging from ass et allocation to risk management. For more than a decade now the econometric literature has made a number of significant leaps forward in the estimation of covariance matrices using financial high-frequency data. This new generation of estimators, commonly referred to as realized covariance estimators, measure precisely the daily covariance of log-prices using intra-daily price information. The literature has proposed an extensive number of procedures that allow us to estimate the covariance efficiently under general assumptions, such as the presence of market microstructure noise and asynchronous trading in the data-generating process (DGP).

Despite the significant leaps forward, the estimation of large realized covariance matrices has a number of hurdles. First, as it has been put forward by Hautsch, Kyj, and Oomen (2012) and Hautsch, Kyj, and Malec (2015), it is hard to estimate precisely the covariance matrix when the number of assets is large. Second, in large systems it is challenging to synthesize effectively the information contained in the covariance matrix and unveil the cross-sectional dependence structure of the data. In this work we propose a realized covariance estimation strategy that tackles simultaneously both of these challenges. The estimation approach consists of using LASSO-type shrinkage to regularize realized covariance estimators. The LASSO procedure detects and estimates the nonzero partial correlations among the daily log-prices. The set of nonzero partial correlations can then be represented as a network. Our proposed estimation approach has different highlights. If the partial correlation structure of the daily log-prices is sufficiently sparse, then the regularized estimator

can deliver substantial accuracy gains over its unregularized counterpart. Moreover, detecting the network of interconnections among the daily log-prices is interesting in the light of the recent strand of research on networks in economics by, among others, Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012), which shows that in highly interconnected systems the most highly interconnected entities influence the aggregate behavior of the entire system.

In its more general version, the framework we work in makes a number of fairly common assumptions on the dynamics of the asset prices (cf. Aït-Sahalia, Mykland, & Zhang, 2005; Bandi & Russell, 2006; Fan, Li, & Yu, 2012). We assume that observed log-prices are equal to the efficient log-prices, which are Brownian semi-martingales, plus a noise term that is due to market microstructure frictions. Prices are observed according to the realization of a counting process driving the arrival of trades/quotes of each asset and are allowed to be asynchronous. The target estimation of interest is an integrated covariance matrix of the efficient daily log-prices.

We introduce a network definition built upon the integrated covariance, which we call the integrated partial correlation network. Assets i and j are connected in the integrated partial correlation network if and only if the partial correlation between i and j implied by the integrated covariance is nonzero. As is well known, the network is entirely characterized by the inverse of the integrated covariance matrix, which we call the integrated concentration matrix. In fact, it has been known since at least Dempster (1972) that if the (i,j)th entry of the inverse covariance matrix is zero, then variables i and j are partially uncorrelated—that is, are uncorrelated conditional on all other assets. Thus the sparsity structure of the integrated concentration matrix determines the partial correlation network dependence structure among the daily log-prices.

We use LASSO to obtain a sparse integrated concentration matrix estimator. The procedure consists of regularizing a consistent realized covariance estimator. Several realized covariance estimators have been introduced in the literature in the presence of market microstructure effects and asynchronous trading. In this work we focus in particular on the two-scales realized covariance estimators (TSRC) and the multivariate (generalized flat-top) realized kernel (MRK) based on pairwise refresh sampling (Aït-Sahalia et al., 2005; Barndorff-Nielsen, Hansen, Lunde, & Shephard, 2011; Fan et al., 2012; Varneskov, 2016). These estimators are then regularized using the GLASSO Friedman, Hastie, and Tibshirani (2011), which shrinks the off-diagonal elements of the inverse of the covariance estimators entries to zero. The procedure allows us to detect the nonzero linkages of the integrated partial correlation network. Moreover, the sparse integrated concentration matrix estimator can be inverted to obtain an estimator of the integrated covariance.

We study the large-sample properties of the realized network estimator, and establish conditions for consistent estimation of the integrated concentration and consistent selection of the integrated partial correlation network. We develop the theory for the TSRC and MRK estimators based on pairwise refresh-sampling built upon the general asymptotic theory developed by Ravikumar, Wainwright, Raskutti, and Yu (2011). The MRK estimator results are obtained by developing a novel concentration inequality, while for the TSRC estimator we apply a concentration inequality derived in Fan et al. (2012). Results are established in a high-dimensional setting; that is, we allow for the total number of parameters to be larger than the number of observations available, to the extent that the proportion of nonzero parameters is small relative to the total. Other realized covariance estimators satisfying an appropriate concentration assumption lead to regularized estimators with similar properties.

A simulation study is used to investigate the finite-sample properties of the procedure. Different specifications of the integrated covariance matrix of the efficient price process are used to assess the precision of the realized network estimator. The procedure is also benchmarked against a set of alternative regularization techniques proposed in the literature, including shrinkage (Ledoit & Wolf, 2004) and factor-based approaches. Among others results, simulations show that when the integrated concentration matrix is indeed sparse the realized network achieves the best performance among the set of candidate regularization procedures we consider.

We apply the realized network methodology to analyze the network structure of a panel of US blue chip stocks throughout 2009 using the TSRC, MRK, as well as the classic realized covariance (RC) estimators. More precisely, we use the realized network to regularize what we call idiosyncratic realized covariance matrix—that is, the residual covariance matrix of the assets after netting out the influence of the market factor. Results show that after controlling for the market factor assets still exhibit a significant amount of cross-sectional dependence. The estimated networks are indeed sparse, with the number of estimated links being roughly 5% of the total possible number of linkages. The distribution of the connections of the assets exhibits power law behavior; that is, the number of connections is heterogeneous and the most interconnected stocks have a large number of connections relative to the total number of links. The stocks in the industrial and energy sectors show a high degree of sectoral clustering; that is, there is a large number of connections among firms in these industry groups. Technology companies, and Google in particular, are the most highly interconnected firms throughout the year. We investigate the usefulness of our procedure from a forecasting perspective by carrying out

a Markowitz-type global minimum variance (GMV) portfolio prediction exercise. We run a horse race among different (regularized) covariance estimators to assess which estimator produces GMV portfolio weights that deliver the minimum out-of-sample GMV portfolio variance. Results show that the realized network significantly improves prediction accuracy irrespective of the covariance estimator used.

We build upon the literature on realized volatility and realized covariance estimation. Important contributions in this area include the work of Andersen, Bollerslev, Diebold, and Labys (2003), Barndorff-Nielsen and Shephard (2004), Aït-Sahalia et al. (2005), Bandi and Russell (2006); Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008), Barndorff-Nielsen et al. (2011), Zhang (2011), and Fan et al. (2012). More precisely, our work is related to the strand of the literature concerned with the estimation and regularization of possibly large realized covariances. Important research in this field includes Wang and Zou (2010), Hautsch et al. (2012; 2015), Tao, Wang, and Zhou (2013); Corsi, Peluso, and Audrino (2014); Kim, Wang, and Zou (2016), and Lunde, Shephard, and Sheppard (2016). This paper also relates the network modeling literature in statistics and econometrics, which includes Meinshausen and Bühlmann (2006), Diebold and Yilmaz (2014), Hautsch, Schaumburg, and Schienle (2014a; 2014b), Barigozzi and Brownlees (2013), and Banerjee and Ghaoui (2008). Last, this paper is related to the literature on covariance matrix regularization. Contributions in this area include the work of Ledoit and Wolf (2004; 2012), Fan, Liao, and Mincheva (2011; 2013). Pourahmadi (2013) provides an introduction to high-dimensional covariance regularization, which includes several of the recent developments of the area.

It is important to highlight the differences between this work and the contributions of Wang and Zou (2010), Tao et al. (2013), and Kim et al. (2016). These papers propose realized covariance regularization procedures based on the assumption that the integrated covariance is sparse, whereas in this paper we impose sparsity assumptions on its inverse. Note that in our framework the integrated covariance matrix is allowed to be nonsparse. Second, the aforementioned contributions are based on thresholding whereas in this paper we rely on LASSO regularization. Last, the LASSO technique used in this work allows us to recover the partial correlation structure of the log-prices, which may give insights into the dependence structure of the assets in the panel.

The rest of the paper is structured as follows. In Section 2 we introduce the base framework and the realized network estimator. The theoretical properties of the estimation procedure are analyzed in Section 3. Section 4 introduces a number of important extensions to the baseline framework. Section 5 contains a simulation exercise to study the properties of the realized network estimator. Section 6 presents an application to a panel of US blue chip stocks. Concluding remarks follow in Section 7.

2 | METHODOLOGY

In this section we introduce the baseline framework and estimation approach. Important extensions of the baseline methodology, including allowing for market microstructure frictions, are considered later in Section 4.

2.1 | Model

Let $y(t) = (y_1(t), \dots, y_n(t))$ denote the n-dimensional log-price vector of n assets at time $t \in [0,1]$. We assume that the dynamics of $y(t)$ are given by

$$y(t) = \int_0^t b(u)\,\mathrm{d}u + \int_0^t \Theta(u)\,\mathrm{d}B(u), \quad t \in [0,1], \tag{1}$$

where $B(t)$ is an n-dimensional Brownian motion. We assume that $y(t)$ is defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,1]}, \mathrm{P})$, where $\mathcal{F}_t \subseteq \mathcal{F}$ is an increasing family of σ-fields satisfying P-completeness and right continuity. We also denote by $\mathcal{F}_t^{(i,j)}$ the restriction of $\mathcal{F}_t$ excluding the ith and jth log-prices. The drift $b(t)$ is an n-dimensional $\mathcal{F}_t$-predictable process, and the spot covolatility process $\Theta(t)$ is an $n \times n$ positive definite random matrix, whose entries are $\mathcal{F}_t$-adapted and càdlàg. Both processes $b(t)$ and $\Theta(t)$ are assumed to be uniformly bounded on $[0,1]$. Similarly to other papers on realized covariance estimation, we do not consider jumps in the $y(t)$ process and we leave this important development for future work.

We consider the $y(t)$ process over a fixed time interval of length 1, which typically represents a day, and we set $y = y(1)$. One of the main estimands of interest in this work is the quadratic covariation matrix of y, that is

$$\Sigma^* = \int_0^1 \Sigma(t)\mathrm{d}t = (\sigma^*_{ij}), \tag{2}$$

where $\Sigma(t) = \Theta(t)\Theta(t)' = (\sigma_{ij}(t))$ is the spot covariance matrix. Throughout the paper we refer to Σ^* as the integrated covariance matrix.

In this work we introduce a network definition for y based on the partial correlation structure implied by the integrated covariance matrix Σ^*. We define the partial correlation ρ^{ij} between y_i and y_j (conditional on $\mathcal{F}_1^{(i,j)}$) as the correlation between ϵ_i and ϵ_j, where ϵ_i and ϵ_j are the prediction errors of the best linear predictors for, respectively, y_i and y_j based on $\{y_k : 1 \leq k \leq n, k \neq i, j\}$ (Peng, Wang, Zhou, & Zhu, 2009). It is well known that if y has covariance Σ^* then we have (Dempster, 1972; Pourahmadi, 2013)

$$\rho^{ij} = \frac{-k^*_{ij}}{\sqrt{k^*_{ii}k^*_{jj}}},$$

where k^*_{ij} denotes the (i,j)-element of the inverse integrated covariance matrix $\mathbf{K}^* = (\Sigma^*)^{-1}$, which we call hereafter the integrated concentration matrix. Partial correlation is one of the standard measures of dependence used to define networks in the literature (Meinshausen & Bühlmann, 2006; Peng et al., 2009). In this paper a network is defined as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices $\mathcal{V} = \{1, 2, \cdots, n\}$ and $\mathcal{E}$ is the set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. In particular, we define the network for y as an undirected graph where the set of vertices corresponds to the set of assets and a pair of assets is connected by an edge iff the corresponding partial correlation is nonzero; that is, $\mathcal{E} = \{(i,j) \in \mathcal{V} \times \mathcal{V}, k^*_{ij} \neq 0, i \neq j\}$. We call this network the integrated partial correlation network. Note that our network definition synthesizes the partial correlation structure of the daily returns of the assets in the panel. It is important to emphasize that the integrated partial correlation network definition captures a particular correlation relation among the daily log-prices. Obviously enough, the absence of correlation between the log-daily prices of two assets does not necessarily imply that the spot prices are also uncorrelated.

2.2 | Estimation

We are interested in (i) estimating the integrated covariance and concentration matrices of the daily log-prices, and (ii) detecting the nonzero entries of the integrated concentration matrix. The estimation strategy we follow consists of applying LASSO-type regularization on the standard realized covariance estimators proposed in the literature.

We assume that the log-prices $y_i(t)$ of all assets $i = 1, \ldots, n$, are discretely observed at a same time grid $T = \{t_1, t_2, \ldots, t_m\}$ where $t_0 = 0 < t_1 < \cdots < t_m = 1$. We consider a generic estimator of the integrated covariance Σ^* denoted $\overline{\Sigma} = (\overline{\sigma}_{ij})$ based on the observations $y_i(t_\ell)$, $i = 1, \ldots, n$, $\ell = 1, \ldots, m$. We assume that this estimator satisfies the following concentration inequality.

Assumption 1. There exist positive constants a_1, a_2 and a_3 such that for all $i, j \in \{1, \ldots, n\}, x \in [0, a_1]$, and m large:

$$\mathbf{P}\left(\left|\overline{\sigma}_{ij} - \sigma^*_{ij}\right| > x | \mathcal{F}_1\right) \leq a_2 m^{\alpha_0} \exp(-a_3 (m^\beta x)^a), \tag{3}$$

for some positive exponents β, a and $\alpha_0 \in \{0, 1\}$.

A natural estimator of the integrated covariance of y in this setting is the so-called realized covariance (RC) estimator. This estimator is the multivariate extension of the realized variance, whose working mechanism is that the quadratic variation of the univariate price process can be approximated by the sum of squared returns over small intervals.

2.2.1 | Realized covariance estimator

The realized covariance estimator $\overline{\Sigma}_{\text{RC}}$ is defined as

$$\overline{\sigma}_{\text{RC},ij} = \sum_{k=1}^{m} (y_{ik} - y_{ik-1})\left(y_{jk} - y_{jk-1}\right),$$

where $y_{ik} = y_i(t_k)$.

Assume that the time grid satisfies

$$\sup_{\ell \in \{1, \cdots, m\}} |t_\ell - t_{\ell-1}| \leq \frac{\kappa}{m}, \tag{4}$$

for some constant $\kappa > 0$. Then Barndorff-Nielsen and Shephard (2004) show that the difference between $\overline{\sigma}_{\text{RC},ij}$ and σ^*_{ij} is asymptotically mixed normal with mean zero and variance of order $O\left(m^{-1}\right)$. Also, it is proved in Fan et al. (2012, Lemma 3) that under the conditions of this section the estimator satisfies Assumption 1 with $\alpha_0 = 0, a = 2$ and $\beta = 12$.

Given an estimator of the integrated covariance $\overline{\Sigma}$ satisfying Assumption 1, we use the graphical LASSO procedure (GLASSO) to estimate the integrated concentration matrix $\mathbf{K}^*$.

2.2.2 | Realized network estimator

Let $\overline{\Sigma}$ be an estimator of the integrated covariance, then we define the realized network estimator of the integrated concentration matrix as

$$\widehat{\mathbf{K}}_\lambda = \underset{\mathbf{K}\in\mathcal{S}^n}{\arg\min} \left\{ \operatorname{tr}(\overline{\Sigma}\mathbf{K}) - \log\det(\mathbf{K}) + \lambda \sum_{i\neq j} |k_{ij}| \right\}, \tag{5}$$

where $\lambda \geq 0$ is the GLASSO tuning parameter and $\mathcal{S}^n$ is the set of $n \times n$ symmetric positive definite matrices. The entries of $\widehat{\mathbf{K}}_\lambda$ are denoted by $(\widehat{k}_{\lambda\, ij})$. The corresponding realized covariance estimator based on the realized network is $\widehat{\Sigma}_\lambda = \widehat{\mathbf{K}}_\lambda^{-1}$.

Observe that Equation 5 defines a shrinkage type estimator. If we set $\lambda = 0$ in Equation 5, we obtain the normal log-likelihood function of the covariance matrix, which is minimized by the inverse realized covariance estimator $(\overline{\Sigma})^{-1}$. If λ is positive, Equation 5 becomes a penalized likelihood function with penalty equal to the sum of the absolute values of the nondiagonal entries in the estimator. The important feature of the absolute value penalty is that, for $\lambda > 0$, some of the entries of the realized network estimator are going to be set to zero. The highlight of this estimator is that it simultaneously estimates and selects the nonzero entries of $\mathbf{K}^*$, thus providing an estimate of the linkages in the network. For this reason we dub the estimator the realized network estimator. Banerjee and Ghaoui (2008) show that the optimization problem in Equation 5 can be solved through a series of LASSO regressions, which motivates an iterative algorithm to solve Equation 5 given in Friedman, Hastle, Hofling, and Tibshirani (2007). For completeness, we provide a description of the algorithm in the Supporting Information Appendix. The highlight of the procedure is that it is straightforward to carry out the minimization of Equation 5 even when the number of series is large. Importantly, the algorithm is also guaranteed to provide a positive definite estimate of the concentration matrix provided that the initial value of the algorithm is a positive definite matrix. Moreover, the algorithm only requires the $\overline{\Sigma}$ estimator to be positive semidefinite (provided that λ is larger than zero). In order to apply the estimator in empirical applications we need to use a selection criterion to pick the value of the tuning parameter λ. In this work we resort to a Bayesian information criterion (BIC)-type criterion defined as

$$\mathrm{BIC}(\lambda) = m \times \left[-\log\det\widehat{\mathbf{K}}_\lambda + \operatorname{tr}\left(\widehat{\mathbf{K}}_\lambda \overline{\Sigma}\right)\right] + \log m \times \#\{(i,j) : 1 \leq i \leq j \leq n, \hat{k}_{\lambda\, ij} \neq 0\},$$

as suggested in Yuan and Lin (2007), among others.

3 | THEORY

In this section, we apply the theory of Ravikumar et al. (2011) to our particular case of an exponential concentration inequality to establish the large sample properties of the realized network estimator defined in Equation 5.

In order to state the results we need to adopt the following notations. Given a matrix $\mathbf{U} = (u_{ij}) \in \mathbb{R}^{\ell\times m}$, we set $||\mathbf{U}||_\infty$, $||\mathbf{U}||_1$, and $|||\mathbf{U}|||_\infty$ to denote $\max_{i,j}|u_{ij}|$, $\sum_{i,j}|u_{ij}|$, and $\max_j \sum_{k=1}^m |u_{jk}|$, where $i \in \{1, 2, \dots, \ell\}$ and $j \in \{1, 2, \dots, m\}$. If $\mathbf{A} = (a_{ij})$ is a $p \times q$ matrix and $\mathbf{B}$ is an $m \times n$ matrix, the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$ is the $pm \times qn$ matrix given by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1q}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p1}\mathbf{B} & \cdots & a_{pq}\mathbf{B} \end{bmatrix}.$$

We index the pm rows of $\mathbf{A} \otimes \mathbf{B}$ by

$$\mathcal{R} = \{(1,1),(2,1),\cdots,(m,1),(1,2),(2,2),\cdots,(m,2),\cdots,(1,p),\cdots,(m,p)\}$$

and the qn columns by

$$\mathcal{C} = \{(1,1),(2,1),\cdots,(n,1),(1,2),(2,2),\cdots,(n,2),\cdots,(1,q),\cdots,(n,q)\}.$$

For any two subsets $\bar{\mathcal{R}} \subset \mathcal{R}$ and $\bar{\mathcal{C}} \subset \mathcal{C}$, we denote by $(\mathbf{A}\otimes\mathbf{B})_{\bar{\mathcal{R}}\bar{\mathcal{C}}}$ the matrix such that $(\mathbf{A}\otimes\mathbf{B})_{(i,j)(c,d)}$ is an entry of $(\mathbf{A}\otimes\mathbf{B})_{\bar{\mathcal{R}}\bar{\mathcal{C}}}$ iff $(i,j) \in \bar{\mathcal{R}}$ and $(c,d) \in \bar{\mathcal{C}}$.

Assumption 2. Consider the $n^2 \times n^2$ matrix $\Gamma^* = \Sigma^* \otimes \Sigma^*$. There exists some $\alpha \in (0,1]$ such that

$$\max_{e\in\mathcal{S}^c} ||\Gamma^*_{e\mathcal{S}}\left(\Gamma^*_{\mathcal{S}\mathcal{S}}\right)^{-1}||_1 \leq 1-\alpha,$$

where $\mathcal{S} = \mathcal{E} \cup \{(i,i)|i \in \mathcal{V}\}$ and $\mathcal{S}^c = \left\{(i,j) \in \mathcal{V}\times\mathcal{V}, k^*_{ij} = 0\right\}$.

Assumption 2 limits the amount of dependence between the nonedge terms (indexed by $\mathcal{S}^c$) and the edge-based terms (indexed by $\mathcal{S}$). The limit is controlled by α: The bigger the α, the smaller is the dependence. In other words, if we set

$$X_{(j,k)} = y_j y_k - E\left(y_j y_k\right), \quad \text{for all } j,k \in \mathcal{V},$$

then the correlation between $X_{(j,k)}$ and $X_{(\ell,m)}$ is low for any $(j,k) \in \mathcal{S}$ and $(\ell,m) \in \mathcal{S}^c$.

In the following, Theorem 2 shows (a) the rate at which the realized network estimator converges to the true value as the sample size m increases, and (b) a lower bound on the probability of correctly detecting the nonzero partial correlations (as well as their signs) as a function of the sample size m. In particular, the estimator is model selection consistent with high probability, when n is large.

Theorem 1. *Assume Assumptions 1 and 2 hold, and choose* $\lambda = \frac{8}{\alpha} m^{-\beta}\left(\frac{\log(a_2 m^{\alpha_0} n^{\tau})}{a_3}\right)^{\frac{1}{\alpha}}$ *in Equation 5, where* $\tau > 2$ *is arbitrary.*

a. *Assume that*

$$m > \left\{\frac{2^{\alpha_0}}{a_3}\log\left[a_2 n^{\tau}\left(a_3^{-\frac{1}{\alpha\beta}} c_0^{\frac{1}{\beta}}\right)^{\alpha_0}\right]\right\}^{\frac{1}{\alpha\beta}} c_0^{\frac{1}{\beta}}, \tag{6}$$

where

$$c_0 := \max\left[\frac{1}{a_1}, 6(1+8\alpha^{-1})^2 d \max\left(C_{\Sigma^*} C_{\Gamma^*}, C_{\Sigma^*}^3 C_{\Gamma^*}^2\right), \frac{a_3^{\frac{1}{\alpha}}}{a_2^{\beta}} \exp\left(\frac{2}{a^2\beta}\right) \mathbf{1}_{\{\alpha_0=1\}}, \frac{1}{\sigma_n}\right]. \tag{7}$$

Here, $\sigma_n = \min_i \sigma_{ii}^*$, *d is the maximum degree of the network—that is, the maximum number of edges that include a vertex—and we have set* $C_{\Gamma^*} = |||(\Gamma_{SS}^*)^{-1}|||_\infty$ *and* $C_{\Sigma^*} = |||\Sigma^*|||_\infty$.

Then,

$$P\left[||\widehat{\mathbf{K}}_\lambda - \mathbf{K}^*||_\infty \le 2\left(1+8\alpha^{-1}\right) C_{\Gamma^*} m^{-\beta}\left(\frac{\log(a_2 m^{\alpha_0} n^{\tau})}{a_3}\right)^{\frac{1}{\alpha}} | \mathcal{F}_1\right] \ge 1 - \frac{1}{n^{\tau-2}}. \tag{8}$$

b. *Define* $\bar{c}_0 = \max\left(c_0, \frac{2C_{\Gamma^*}(1+8\alpha^{-1})}{k_n}\right)$, *where* k_n *is the minimum absolute value of the nonzero entries of* $\mathbf{K}^*$. *Assume that*

$$m > \left\{\frac{2^{\alpha_0}}{a_3}\log\left[a_2 n^{\tau}\left(a_3^{-\frac{1}{\alpha\beta}} \bar{c}_0^{\frac{1}{\beta}}\right)^{\alpha_0}\right]\right\}^{\frac{1}{\alpha\beta}} \bar{c}_0^{\frac{1}{\beta}}.$$

Then,

$$\mathrm{P}\left[sign(\widehat{k}_{\lambda ij}) = sign(k_{ij}^*),\ \forall i,j \in \mathcal{V} | \mathcal{F}_1\right] \ge 1 - \frac{1}{n^{\tau-2}}.$$

Let us give the intuition behind the proof of this result. The assumption in Equation 6 implies that m is sufficiently large, so that the estimation error $|\overline{\sigma}_{ii} - \sigma_{ii}^*|$ is not larger than $\sigma_n = \min_i \sigma_{ii}^*$ with high probability for all i and, consequently, $\overline{\Sigma}$ will have positive diagonal entries. In this case, the optimization problem of Equation 5 is convex and has a unique solution. Let $\widetilde{\mathbf{K}}_\lambda = (\tilde{k}_{\lambda ij})$ be the solution of Equation 5 under the constraint $\tilde{k}_{\lambda ij} = 0$ if $k_{ij}^* = 0$ (see Equation A-1 in the Supporting Information Appendix). Based on the primal–dual witness construction (see, e.g., Ravikumar et al., 2011), we have that $\widetilde{\mathbf{K}}_\lambda = \widehat{\mathbf{K}}_\lambda$ when $||\widetilde{\mathbf{K}}_\lambda - \mathbf{K}^*||_\infty$ and $||\overline{\Sigma} - \Sigma^*||_\infty$ are not larger than an appropriately defined constant. After straightforward computations we show that for an appropriate choice of λ and m large enough we have that $||\overline{\Sigma} - \Sigma^*||_\infty$ and $||\widetilde{\mathbf{K}}_\lambda - \mathbf{K}^*||_\infty$ satisfy this condition. Therefore, with high probability we have that $\widetilde{\mathbf{K}}_\lambda = \widehat{\mathbf{K}}_\lambda$. In part (b) of Theorem 1 we introduce the parameter k_n, which is a lower bound on the smallest nonzero entries of $\mathbf{K}^*$ in absolute value. We then show that for m sufficiently large we have that with high probability $|\widehat{k}_{\lambda ij} - k_{ij}^*|$ is smaller than k_n. This in turn is used to establish that the signs of $\widehat{k}_{\lambda ij}$ and k_{ij}^* are equal. Observe that when Assumption 1 holds with $\alpha_0 = 0$, Theorem 1 is a direct application of Theorems 1 and 2 of Ravikumar et al. (2011). We give the proof of Theorem 1 for the case $\alpha_0 = 1$ in the Supporting Information Appendix.

The parameter d, the maximum degree in the network, determines the precision of the estimator. It ranges from 0 (empty network) to n. To explore its effects, let us assume that the parameters C_{Σ^*}, C_{Γ^*}, α and σ_n in Theorem 1 remain constant as a function (n,m,d). In this case, when $d = 0$, Condition 6 means that m should not be smaller than $O\left((\log n)^{\frac{1}{\alpha\beta}}\right)$. When $d = n$, Condition 6 means that m should not be smaller than $O\left((\log n)^{\frac{1}{\alpha\beta}} n^{\frac{1}{\beta}}\right)$, since in this case $c_0 = O(n)$. In other words, the more sparse the network is, the fewer observations are required to estimate the concentration matrix accurately.

4 | EXTENSIONS

4.1 | Microstructure noise and asynchronicity

Rather than the efficient price, it is customary to assume that the econometrician observes the transaction (or midquote) price. This differs from the efficient price because trades (quotes) are affected by an array of market frictions that go under the umbrella term of market microstructure. Moreover, it is common to assume that the trades (quotes) of different assets are executed (posted) asynchronously. In this section we extend the baseline framework of Section 2 and introduce a number of realized covariance estimators designed to handle microstructure noise and asynchronous trading.

We assume that the log-prices of each asset i are observed asynchronously on different time grids $T_i = \{t_{i1}, \cdots, t_{im_i}\}$, $i = 1, \dots, n$. For each asset $i = 1, \dots, n$ the econometrician observes the transaction (or midquote) prices $x_i(t_{i\ell})$ defined as

$$x_i(t_{i\ell}) = y_i(t_{i\ell}) + u_i(t_{i\ell}), \tag{9}$$

where $u_i(t_{i\ell})$ denotes the microstructure noise associated with the ℓth trade. Precise assumptions on the noise are spelled out in what follows.

A standard technique used to handle asynchronous trading for realized covariance estimation is refresh time sampling, which was introduced by Martens (2004). Several variants of this technique exist, like the pairwise and groupwise refresh time approaches, used in Fan et al. (2012), Lunde et al. (2016) and Hautsch et al. (2012). In this work we use pairwise refresh time sampling. Pairwise refresh time sampling-based covariance estimation consists of estimating each entry of the covariance separately. The i,j-entry of the matrix is computed by first synchronizing the observations of assets i and j using refresh time and then estimating the covariance between assets i and j using the synchronized data. We provide an exact definition of the pairwise refresh time sampling procedure in the Supporting Information Appendix. Note that this approach does not guarantee that the covariance estimator is positive definite, as each covariance entry is estimated using different subsets of observations. Let $x_i^r = \{x_i(t_{i1}^r), \cdots, x_i(t_{im}^r)\}$ and $x_j^r = \{x_j(t_{j1}^r), \cdots, x_j(t_{jm}^r)\}$ denote the pairwise refresh time sampling prices for assets i and j, where t_{ik}^r and t_{jk}^r are the synchronized timestamps. We use the shorthand notation $x_{\ell k}^r$, $y_{\ell k}^r$ and $u_{\ell k}^r$ to denote $x_\ell(t_{\ell k}^r)$, $y_\ell(t_{\ell k}^r)$ and $u_\ell(t_{\ell k}^r)$, respectively. Also, we define M_0 as the minimum pairwise refresh sample size across all pairs of assets.

After the data have been opportunely synchronized, a number of market microstructure noise robust estimators can be applied. In this work we focus on two leading robust estimators proposed in the literature: the two-scales realized covariance estimator (TSRC) and the multivariate (generalized flat-top) realized kernel estimator (MRK).

In this work the noise process $u_i(t_{i\ell})$ is assumed to be independent of $\mathcal{F}_1$ and normally distributed with mean zero and variance η^2. As far as its dependence properties are concerned, we consider two different settings: The propositions for the TSRC estimator assume that each (i,j) refreshed pair (u_{ik}^r, u_{jk}^r) is i.i.d., whereas the propositions for the MRK estimator assume that (u_{ik}^r, u_{jk}^r) is multivariate $\mathcal{M}$-dependent for some constant $\mathcal{M} > 0$.[1]

4.1.1 | Two-scales realized covariance estimator

The two–scales realized covariance estimator (TSRC) proposed in Zhang (2011) is a multivariate extension of the two-scales estimator introduced by Aït-Sahalia et al. (2005). The TSRC estimator $\overline{\Sigma}_{\text{TS}}$ based on pairwise refresh time is defined as

$$\overline{\sigma}_{\text{TS},ij} = \frac{1}{K}\sum_{k=K+1}^{m}\left(x_{ik}^r - x_{ik-K}^r\right)\left(x_{jk}^r - x_{jk-K}^r\right) - \frac{m_K}{m_J}\frac{1}{J}\sum_{k=J+1}^{m}\left(x_{ik}^r - x_{ik-J}^r\right)\left(x_{jk}^r - x_{jk-J}^r\right),$$

where $m_K = \frac{m-K+1}{K}$ and $m_J = \frac{m-J+1}{J}$.

Zhang (2011) shows that the optimal choice of K has order $K = O(m^{\frac{2}{3}})$, and J can be taken to be a constant such as 1. The first component of this estimator is the average of K realized variances, and it converges to σ_{ij}^* in the absence of noise. The second component is set to correct the bias caused by the noise. Under the optimal choice of K and J, the estimation error is asymptotically mixed normal with zero mean and variance of order $O\left(m^{-\frac{1}{3}}\right)$. If we further assume that $\frac{1}{2}m^{\frac{1}{3}} \leq m_K \leq 2m^{\frac{1}{3}}$, then, under the condition that the synchronized observation times satisfy Condition 4, Fan et al. (2012) show that this estimator satisfies Assumption 1 with $\beta = \frac{1}{6}$, $a = 2$ and $\alpha_0 = 0$, and thus Theorem 1. Observe that

[1] In other words, we assume that $u_{ik_1}^r$ is independent of $u_{jk_2}^r$ when $|k_1 - k_2| > \mathcal{M}$.

since M_0 is defined as the minimum pairwise refresh sample size across all pairs of assets, we should replace m with M_0 when applying Theorem 2 to TSRC.

In the simulation and empirical studies, for each pair of assets we choose J as 1 and K as the average number of trades in 1 minute.[2]

4.1.2 | Multivariate generalized flat-top realized kernel estimator

The MRK is proposed in Varneskov (2016), and builds upon the realized kernel estimators introduced in Barndorff-Nielsen et al. (2008) and Barndorff-Nielsen et al. (2011). The MRK estimator $\overline{\Sigma}_{\mathrm{K}}$ based on pairwise refresh time is defined as

$$\overline{\sigma}_{\mathrm{K}ij} = \gamma_0\left(x_i^r, x_j^r\right) + \frac{1}{2}\sum_{h=1}^{H(1+H^{-\nu})} k\left(\frac{h}{H}\right)\left[\gamma_h\left(x_i^r, x_j^r\right) + \gamma_{-h}\left(x_i^r, x_j^r\right) + \gamma_h\left(x_j^r, x_i^r\right) + \gamma_{-h}\left(x_j^r, x_i^r\right)\right], \tag{10}$$

where for each $h \in \{-H(1+H^{-\nu}), \ldots, H(1+H^{-\nu})\}$ we have $\gamma_h(x_i^r, x_j^r) = \sum_{p=1}^{m}(x_{ip}^r - x_{ip-1}^r)(x_{jp-h}^r - x_{jp-h-1}^r)$. The flat-top kernel function k is defined as

$$k(x) = \mathbf{1}_{\{|x|\leq H^{-\nu}\}} + \mu(|x| - H^{-\nu})\mathbf{1}_{\{|x|>H^{-\nu}\}},$$

where $\nu \in (0,1)$ and the function μ satisfies the regularity conditions in Definition 3 of Varneskov (2016). Note that in Equation (10) we are implicitly assuming that $H^{-\nu}H$ is an integer. Under fairly general assumptions on the noise process (which allow, for instance, for α-mixing dependence) Varneskov shows that for a choice of H of $O(m^{\frac{1}{2}})$, the estimation error is asymptotically mixed normal with zero mean and variance of order $O\left(m^{-\frac{1}{2}}\right)$.

In order to apply Theorem 1 to this estimator, we establish an appropriate concentration inequality. We emphasize that our result is proven under different conditions on the kernel function and the noise in comparison to those in Varneskov (2016). As far as the kernel function is concerned we assume that (i) $\mu(0) = 1$; (ii) $\mu(x)$ is twice differentiable with bounded derivatives on $[0,1]$; (iii) $\mu(1) = \mu'(0) = \mu'(1) = 0$.

Theorem 2. *If the synchronized observation times satisfy condition (Equation (4), then there exist positive constants a_1, a_2 and a_3 such that for all $i,j \in \{1, \ldots, n\}$, $x \in [0, a_1]$, and M_0 large,*

$$\mathrm{P}\left(\left|\overline{\sigma}_{\mathrm{K}ij} - \sigma_{ij}^*\right| > x | \mathcal{F}_1\right) \leq a_2 M_0 \exp\left(-a_3 M_0^{1/4} x\right).$$

Therefore, the MRK satisfies Assumption 1 with $\beta = \frac{1}{4}$, $a = 1$, and $\alpha_0 = 1$. Hence, we can apply Theorem 1, and we obtain that the estimation error of $\widehat{\mathbf{K}}_\lambda$ converges to zero at rate $M_0^{-\frac{1}{4}}\sqrt{\log n}$ (assuming that all other parameters including α, d, C_{Γ^*} and C_{Σ^*} are constants). Thus, in this case, $M_0^{\frac{1}{2}}$ is required to be large compared to $\log n$ to make the error small in probability. Note that this result is analogous to that obtained in Tao et al. (2013), where the same convergence rate $M_0^{-\frac{1}{4}}\sqrt{\log n}$ is obtained for a multi-scale realized covariance estimator. Moreover, according to Tao et al., this rate is the optimal rate for the estimation of the integrated covariance matrix when noise is present.

In the simulation and empirical studies, for each pair of assets we follow closely the implementation procedure described in Varneskov (2016). In particular, in our MRK implementation we use the Parzen kernel and we choose the parameters H and ν using the plug-in approach detailed in that paper using refreshed time sampled prices, sampled on average every 20 seconds.

4.2 | Factor structure

Classic asset pricing theory models like the CAPM or APT imply that the unexpected rate of return of risky assets can be expressed as a linear function of few common factors and an idiosyncratic component. Factors induce a fully interconnected partial correlation network structure. In this case, it is natural to carry out network analysis on the partial correlation structure of the assets after netting out the influence of common sources of variation. In this section we propose a modification of our network definition for such systems. Also, we propose a modified covariance estimation strategy analogous to that put forward in Fan, Fan, and Lv (2008) and Fan et al. (2011) that is based on the particular structure of the system.

[2]In the previous version of the paper we choose K as the optimal bandwidth for the realized two-scale volatility estimator of the two assets, following the procedure detailed in Aït-Sahalia et al. (2005). Results are roughly analogous to those presented here.

We augment the y process with additional k components representing factors. The dynamics of the augmented system are assumed to be the same as that described in Equation (1). Moreover, the factors are assumed to be observed, as is commonly done in the empirical finance literature and also as in Fan et al. (2008). The integrated covariance of the augmented system can then be partitioned as an $(n + k) \times (n + k)$ matrix

$$\Sigma^* = \begin{bmatrix} \Sigma^*_{AA} & \Sigma^*_{FA} \\ \Sigma^*_{AF} & \Sigma^*_{FF} \end{bmatrix}, \tag{11}$$

where A and F denote, respectively, the blocks of assets and factors.

The covariance of the assets can be expressed as the sum of the systematic and idiosyncratic components, that is:

$$\Sigma^*_{AA} = \mathbf{B}\Sigma^*_{FF}\mathbf{B}' + \Sigma^*_I,$$

where

$$\mathbf{B} = \Sigma^*_{AF}\left[\Sigma^*_{FF}\right]^{-1} \text{ and } \Sigma^*_I = \Sigma^*_{AA} - \Sigma^*_{AF}\left[\Sigma^*_{FF}\right]^{-1}\Sigma^*_{FA}.$$

If the factors are pervasive ($\mathbf{B}$ is not sparse), then the concentration matrix of the assets cannot be sparse. In these cases, rather than proposing a network definition on the basis of the partial correlations of the system, we propose a network definition based on the idiosyncratic partial correlations—that is, the partial correlations implied by the idiosyncratic covariance matrix Σ^*_I. Precisely, we define the idiosyncratic integrated partial correlation network as the network whose set of edges is given by

$$\mathcal{E}_I = \left\{ (i,j) \in \mathcal{V} \times \mathcal{V}, k^*_{Iij} \neq 0, i \neq j \right\},$$

where k^*_{Iij} is the i,j-entry of the matrix $\mathbf{K}^*_I = (\Sigma^*_I)^{-1}$.

Let $\overline{\Sigma}$ be an appropriate estimator of the integrated covariance of the augmented system and consider partitioning the estimated covariance matrix analogously to Equation (11):

$$\overline{\Sigma} = \begin{bmatrix} \overline{\Sigma}_{AA} & \overline{\Sigma}_{FA} \\ \overline{\Sigma}_{AF} & \overline{\Sigma}_{FF} \end{bmatrix}.$$

Then, a natural estimator of the idiosyncratic realized covariance estimator $\overline{\Sigma}_I = (\overline{\sigma}_{Iij})$ is

$$\overline{\Sigma}_I = (\overline{\sigma}_{Iij}) = \overline{\Sigma}_{AA} - \overline{\Sigma}_{FA}\left[\overline{\Sigma}_{FF}\right]^{-1}\overline{\Sigma}_{AF}. \tag{12}$$

The following corollary establishes the concentration inequality of the estimator $\overline{\Sigma}_I$ using the one for $\overline{\Sigma}$.

Corollary 1. *If Assumption 1 holds, then there exist positive constants b_1, b_2 and b_3 such that for all $i,j \in \{1, \dots, n\}$, $x \in [0, b_1]$, and M_0 large:*

$$\mathrm{P}\left(\left| \overline{\sigma}_{Iij} - \sigma^*_{Iij} \right| > x | \mathcal{F}_1 \right) \leq b_2 M_0^{\alpha_0} \exp(-b_3 (M_0^{\beta} x)^a),$$

where β, a and α_0 are the constants from Assumption 1.

The realized network estimator can thus be applied to regularize the idiosyncratic realized covariance matrix and estimate the idiosyncratic partial correlation network. Moreover, the covariance matrix of the assets can be estimated as

$$\widehat{\Sigma}_{AA} = \overline{\mathbf{B}}\overline{\Sigma}_{FF}\overline{\mathbf{B}}' + \widehat{\Sigma}_{I\lambda},$$

where $\widehat{\Sigma}_{I\lambda}$ denotes the realized covariance estimator implied by the realized network. Note that this estimation strategy is analogous to that proposed in Fan et al. (2011).

5 | SIMULATION STUDY

In this section we carry out a simulation study to assess the finite-sample properties of the realized network estimator. The simulation exercise consists of simulating 1 day of high-frequency data and to apply the realized network estimator to estimate the integrated covariance and the integrated concentration matrices. Different specifications of the covariance matrix of the efficient price process are used to assess the usefulness of the realized network estimator depending on the underlying cross-sectional dependence structure of the data. The realized network estimator is also benchmarked against a set of alternative covariance regularization procedures proposed in the literature.

In our simulation study we employ a DGP analogous to that used in the study of Kim et al. (2016). The efficient log-price is generated according to a zero-drift version of the model in Equation 1, that is:

$$y(t) = \int_0^t \Theta(u)\,\mathrm{d}B(u),$$

where $B(t)$ is an n-dimensional Brownian motion and $\Theta(t)$ is the Cholesky factor of the spot covariance matrix $\Sigma(t)$. In our numerical implementation, a trading day is 6.5 hours long and the simulation of the continuous time process is carried out using the Euler scheme with a discretization step of 5 seconds. In our simulation study we consider different cross-sectional sizes n equal to 25, 50 and 100.

The spot covariance matrix is defined as

$$\Sigma(t) = \kappa(t)\mathbf{V}(t)\mathbf{D}\mathbf{V}(t),$$

where $\kappa(t)$ is a scalar generated as

$$\kappa(t) = \frac{e^{2f(t)} - 1}{e^{2f(t)} + 1}, \quad f(t) = \int_0^t a_\kappa[b_\kappa - f(u)] + c_\kappa f(u)\ \mathrm{d}W_\kappa(u),$$

where $W_\kappa(t)$ is a standard one-dimensional Brownian motion and $(a_\kappa, b_\kappa, c_\kappa)' = (1, 0.4, 0.64)'$; $\mathbf{V}(t)$ is a diagonal $n \times n$ matrix whose ith diagonal entry $v_i(t)$ is

$$v_i(t) = \int_0^t a_v \left[b_v - v_i(u)\right] + c_v\sqrt{v_i(u)}\,\mathrm{d}W_{v_i}(u),$$

where $W_{v_i}(t)$ is a standard one-dimensional Brownian motion and $(a_v, b_v, c_v)' = (0.000042, 0.8, 0.0058)'$; and $\mathbf{D}$ is an $n \times n$ positive definite matrix that determines the cross-sectional dependence structure of the assets and that is specified below. The processes $W_\kappa(t)$, $W_{v_i}(t)$, and $B(t)$ are independent among each other.

Three different specifications of the matrix $\mathbf{D}$ are adopted. In the first simulation design (Design 1), we pick a specification for $\mathbf{D}$ which induces a sparse partial correlation structure among the assets in the panel. In particular, we choose $\mathbf{D}$ as a function of a realization of an Erdös–Renyi random graph. The Erdös–Renyi random graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected graph defined over a fixed set of vertices $\mathcal{V} = \{1, \cdots, n\}$ and a random set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, where the existence of an edge between any pair of vertices is determined by an independent Bernoulli trial with probability p. We generate $\mathbf{D}$ by first simulating an Erdös–Renyi random graph $\mathcal{G}$ and then setting $\mathbf{D}$ equal to

$$\mathbf{D} = [\mathbf{I}_n + \mathbf{E} - \mathbf{F}]^{-1},$$

where $\mathbf{I}_n$ is the n-dimensional identity matrix, and $\mathbf{E}$ and $\mathbf{F}$ are respectively the degree matrix and the adjacency matrix of the random graph $\mathcal{G}$. The model for $\mathbf{D}$ is such that the underlying random graph structure determines the sparsity structure of the integrated concentration matrix. Also, note that $\mathbf{D}$ is symmetric positive definite by construction. In the simulation we set p equal to $2/n$. In this scenario (i.e., when np is greater than one) the Erdös–Renyi random graph will almost surely have a giant component—that is, a connected component containing a constant fraction of the entire set of network vertices. Thus the highlight of the model is that the generated concentration matrix is sparse whereas the corresponding covariance matrix is not.

In the second simulation design (Design 2), we pick a specification for $\mathbf{D}$ based on a factor model. We set $\mathbf{D}$ as

$$\mathbf{D} = \mathbf{G}\mathbf{I}_k\mathbf{G}' + \mathbf{I}_n,$$

where $\mathbf{G}$ is an $n \times k$ matrix whose entries are i.i.d. normal Gaussian draws with mean zero and unit variance. In the simulation we set the number of factors k to 2. Note that this scenario is challenging for the realized network estimator in that the inverse covariance matrix implied by the model is not sparse.

Last, in the third simulation design (Design 3), we set the $\mathbf{D}$ matrix as

$$[\mathbf{D}]_{ij} = \begin{cases} \rho + \rho^{|i-j|} & \text{if } i \neq j \\ 1 & \text{otherwise,} \end{cases}$$

and set ρ equal to 0.25. Note that also in this scenario the inverse covariance matrix is not sparse and the covariance matrix does not have a factor representation.

We assume that the econometrician observes asynchronously a contaminated version of the efficient price, as specified in Equation 9. Prices are observed according to the realization of a Poisson process with a constant intensity calibrated to have 20 trades per minute on average. The market microstructure contamination $u_i(t_{i\ell})$ consists of a zero-mean MA(1)

process $u_i(t_{i\ell}) = \varepsilon_i(t_{i\ell}) - 0.5\varepsilon_i(t_{i\ell-1})$ where $\varepsilon_i(t_{i\ell})$ is drawn from a Student t-distribution with 10 degrees of freedom which is further rescaled to have a standard deviation of 0.05.

Different approaches are used to estimate the integrated covariance and integrated concentration matrices. First, we estimate the integrated covariance using the 1-minute frequency RC, the pairwise-refreshed time TSRC and the pairwise-refreshed time MRK. For the RC we employ a 1-minute sampling frequency rather than the classic 5-minute frequency so that we have enough intra-daily observations to obtain a positive definite estimator when the cross-sectional dimension n is 100. The bandwidth parameters of the TSRC and MRK are computed using the plug-in rules previously described. It is important to stress that the TSRC and MRK estimators are not guaranteed to be positive definite. When the estimators are indefinite we apply an eigenvalue cleaning algorithm used in Hautsch et al. (2012) (reported in the Supporting Information Appendix) to obtain a positive definite estimator.[3] For each realized covariance estimator, we consider a number of different regularization procedures. First, we use the realized network estimator defined in Section 2 using the BIC to determine the optimal amount of shrinkage to apply. Note that one of the inputs of the BIC is the number of observations used to compute the estimator. When using pairwise-refresh sampling, however, this number is different for each entry of the covariance matrix. Similarly to Hautsch et al. (2012), we opt for a conservative choice of this quantity, and we set it to the minimum number of refresh time observations across all pairs. The next regularized estimator we consider is based on a factor approximation of the covariance matrix (Fan et al., 2013). It is defined as

$$\widehat{\Sigma}_F = \sum_{i=1}^{k} \hat{\xi}_i \hat{e}_i \hat{e}_i' + \widehat{\mathbf{R}}_k, \tag{13}$$

where $\hat{\xi}_i$ and $\hat{e}_i$ denote the eigenvalues (in increasing order) and corresponding eigenvectors obtained from the spectral decomposition of the unregularized realized estimator $\overline{\Sigma}$, and $\widehat{\mathbf{R}}_k$ is $\mathrm{diag}(\overline{\Sigma} - \sum_{i=1}^{k} \hat{\xi}_i \hat{e}_i \hat{e}_i')$. Last, we consider the shrinkage estimator proposed by Ledoit and Wolf (2004). Let $\overline{\Sigma}$ denote the unregularized realized covariance (computed using either the RC, TSRC or MRK estimators). The shrinkage estimator is defined as

$$\widehat{\Sigma}_{\mathrm{LW}} = \alpha_1 \mathbf{I}_n + \alpha_2 \overline{\Sigma}, \tag{14}$$

where α_1 and α_2 are two tuning parameters chosen to minimize the risk of the estimator that we set following Ledoit and Wolf (2004) and Hautsch et al. (2015).[4]

Different metrics can be used to evaluate the performance of covariance estimators (Laurent, Rombouts, & Violante, 2011). Here we rely on two classic loss functions: The Kullback–Leibler and root mean square error (RMSE). A classic loss function used for the evaluation of covariance matrix estimators is the Kullback–Leibler loss proposed by Stein (1956) and Pourahmadi (2013), which is defined as

$$\mathrm{KL}\left(\widehat{\Sigma}\right) = \mathrm{tr}(\widehat{\Sigma}\mathbf{K}^*) - \log|\widehat{\Sigma}\mathbf{K}^*| - n.$$

Following Hautsch et al. (2012), we also consider an RMSE-type loss based on the Frobenius norm of the covariance matrix, which is defined as

$$\mathrm{RMSE}\left(\widehat{\Sigma}\right) = \sqrt{\sum_{i,j=1}^{n} (\hat{\sigma}_{ij} - \sigma_{ij}^*)^2}.$$

We perform 10,000 Monte Carlo replications of the simulation exercise for each simulation design and report summary statistics on the performance of the estimators in Table 1. The table reports the average of the KL and RMSE losses of the estimators in the three simulation designs.

A number of comments are in order. First, using regularization is always advantageous with respect to the unregularized estimator even when the regularization scheme does not match the characteristics of the DGP. Gains are more pronounced when the dimension of the system is larger. Second, we note that using a regularization technique whose shrinking target is closer to the true underlying structure of the DGP produces the largest gains. In particular, it is easy to see that when the partial correlation structure of the data is sparse (Design 1) the realized network estimator is the

[3]It is important to emphasize that this procedure is designed for the sample covariance estimator of independent data. Nevertheless, we found the performance of this procedure in the version proposed by Hautsch et al. (2012) to work satisfactorily in both the simulations and the empirical application.

[4]More precisely, we compute values of α_1 and α_2 using the formulas provided in Ledoit and Wolf (2004) using the panel of equally spaced differences at the 1-minute frequency.

TABLE 1 Simulation study

		KL				RMSE			
DGP	**Est.**	**No regular.**	**Network**	**Factor**	**Shrinkage**	**No regular.**	**Network**	**Factor**	**Shrinkage**
$n = 25$									
D1	RC	2.867	2.207	2.355	2.472	1.048	0.794	0.937	0.986
	TSRC	1.947	1.874	2.537	2.416	1.055	0.816	0.898	1.041
	MRK	1.594	1.257	1.552	1.482	0.857	0.647	0.621	0.595
D2	RC	2.364	0.601	0.335	0.727	1.098	0.915	0.728	1.145
	TSRC	1.783	0.421	0.577	0.740	1.310	1.155	0.727	1.451
	MRK	1.553	0.427	0.328	0.711	0.899	0.711	0.582	0.895
D3	RC	2.374	0.443	0.674	0.286	1.043	0.969	1.130	0.992
	TSRC	1.594	0.568	0.936	0.374	1.040	1.089	1.015	0.977
	MRK	1.463	0.547	0.530	0.309	0.864	0.807	0.784	0.742
$n = 50$									
D1	RC	5.330	3.924	4.668	5.245	2.061	1.216	1.450	1.867
	TSRC	3.886	3.307	3.791	3.610	1.947	1.158	1.399	1.895
	MRK	3.738	2.401	3.241	3.554	1.708	1.127	0.957	1.121
D2	RC	4.865	1.360	0.537	0.839	2.156	1.703	1.330	2.255
	TSRC	3.568	1.144	0.852	0.964	2.372	2.012	1.340	2.652
	MRK	3.542	1.226	0.483	0.963	1.742	1.324	1.059	1.725
D3	RC	5.055	1.643	1.789	0.511	2.058	1.857	1.515	1.154
	TSRC	3.095	1.947	2.536	0.629	1.948	1.772	1.372	1.210
	MRK	3.263	2.140	1.435	0.579	1.697	1.567	1.102	0.864
$n = 100$									
D1	RC	23.638	8.588	10.389	11.389	4.080	2.125	2.184	2.587
	TSRC	22.396	7.372	10.762	11.586	3.720	1.953	2.069	2.628
	MRK	24.570	5.877	6.410	6.834	3.381	1.141	1.413	1.552
D2	RC	22.490	3.515	0.803	3.495	4.306	3.379	2.428	4.469
	TSRC	18.183	2.960	1.232	4.699	4.897	4.146	2.709	4.394
	MRK	21.723	3.492	0.732	5.054	3.554	2.716	2.022	3.529
D3	RC	24.061	6.609	2.776	1.969	4.096	3.657	2.297	2.121
	TSRC	13.599	7.281	3.899	2.286	3.780	3.379	2.348	1.963
	MRK	16.971	9.038	2.245	2.218	3.395	3.129	1.713	1.645

Note. The table reports the KL and RMSE average losses of the unregularized RC, TSRC, and MRK estimators (No regular) as well as their regularized versions (Shrinkage, Factor, Network) in the three simulation designs of Section 5.

best-performing regularization technique. Analogously, the factor-based regularization performs best in Design 2 and shrinkage regularization in Design 3. It is also interesting to report how many times we resort to the eigenvalue cleaning procedure for the TSRC and MRK estimators: In our simulations we had to apply the cleaning step on average 84.3% of the times for the TSRC and 72.8% for the MRK estimator. Overall, results convey that the gains by using the realized network estimator when the partial correlation structure is sparse can be substantial when the system is large.

6 | EMPIRICAL APPLICATION

We use the realized network estimator to analyze the dependence structure of a panel of US blue chip stocks from the NYSE throughout 2009. We then engage in a Markowitz-style global minimum variance portfolio prediction exercise.

6.1 | Data and estimation

We consider a sample of 100 liquid US blue chip stocks that have been part of the S&P 100 index for most of the 2000s. We also include in the panel the SPY ETF: the ETF tracking the S&P 500 index. We work with tick-by-tick transaction prices obtained from the NYSE-TAQ database. Before proceeding with the econometric analysis, the data are filtered using standard techniques described in Brownlees and Gallo (2006) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009). The full list of tickers, company names and industry groups is reported in Table 2.

TABLE 2 Tickers, company names and sectors

Ticker symbol	Company name	GICS sector	Ticker symbol	Company name	GICS sector
AMZN	Amazon.com	Consumer discretionary	ABT	Abbott Laboratories	Healthcare
CMCSA	Comcast	Consumer discretionary	AMGN	Amgen	Healthcare
DIS	Walt Disney	Consumer discretionary	BAX	Baxter International	Healthcare
F	Ford Motor	Consumer discretionary	BMY	Bristol-Myers Squibb	Healthcare
FOXA	Twenty-First Century Fox	Consumer discretionary	GILD	Gilead Sciences	Healthcare
GM	General Motors	Consumer discretionary	JNJ	Johnson & Johnson	Healthcare
HD	Home Depot	Consumer discretionary	LLY	Lilly & Co.	Healthcare
LOW	Lowes	Consumer discretionary	MDT	Medtronic	Healthcare
MCD	McDonalds	Consumer discretionary	MRK	Merck	Healthcare
NKE	NIKE	Consumer discretionary	PFE	Pfizer	Healthcare
SBUX	Starbucks	Consumer discretionary	UNH	United Health Group	Healthcare
TGT	Target	Consumer discretionary	BA	Boeing Company	Industrials
TWX	Time Warner Inc.	Consumer discretionary	CAT	Caterpillar	Industrials
CL	Colgate-Palmolive	Consumer staples	EMR	Emerson Electric	Industrials
COST	Costco Co.	Consumer staples	FDX	FedEx	Industrials
CVS	CVS Caremark	Consumer staples	GD	General dynamics	Industrials
KO	The Coca Cola Company	Consumer staples	GE	General Electric	Industrials
MDLZ	Mondelez International	Consumer staples	HON	Honeywell Intl	Industrials
MO	Altria Group Inc.	Consumer staples	LMT	Lockheed Martin	Industrials
PEP	PepsiCo Inc.	Consumer staples	MMM	3M Company	Industrials
PG	Procter & Gamble	Consumer staples	NSC	Norfolk Southern	Industrials
PM	Philip Morris	Consumer staples	RTN	Raytheon Co.	Industrials
WAG	Walgreen	Consumer staples	UNP	Union Pacific	Industrials
WMT	Wal-Mart Stores	Consumer staples	UPS	United Parcel Service	Industrials
APA	Apache	Energy	UTX	United Technologies	Industrials
APC	Anadarko Petroleum	Energy	AAPL	Apple	Information technology
COP	ConocoPhillips	Energy	ACN	Accenture	Information technology
CVX	Chevron	Energy	CSCO	Cisco Systems	Information technology
DVN	Devon Energy	Energy	EBAY	eBay	Information technology
HAL	Halliburton Co.	Energy	EMC	EMC	Information technology
NOV	National Oilwell Varco	Energy	FB	Facebook	Information technology
OXY	Occidental Petroleum	Energy	GOOG	Google Inc.	Information technology
SLB	Schlumberger Ltd.	Energy	HPQ	Hewlett-Packard	Information technology
XOM	Exxon Mobil	Energy	IBM	IBM	Information technology
AIG	AIG	Financials	INTC	Intel	Information technology
ALL	Allstate	Financials	MA	Mastercard	Information technology
AXP	American Express Co.	Financials	MSFT	Microsoft	Information technology
BAC	Bank of America	Financials	ORCL	Oracle	Information technology
BK	Bank of New York	Financials	QCOM	QUALCOMM Inc.	Information technology
BRK.B	Berkshire Hathaway	Financials	TXN	Texas Instruments	Information technology
C	Citigroup Inc.	Financials	V	Visa Inc.	Information technology
COF	Capital One Financial	Financials	DD	Du Pont	Materials
GS	Goldman Sachs Group	Financials	DOW	Dow Chemical	Materials
JPM	JPMorgan Chase & Co.	Financials	FCX	Freeport-McMoran	Materials
MET	MetLife Inc.	Financials	MON	Monsanto Co.	Materials
MS	Morgan Stanley	Financials	T	AT&T Inc.	Telecommunications
SPG	Simon Property Group	Financials	VZ	Verizon Communications	Telecommunications
USB	U.S. Bancorp	Financials	AEP	American Electric Power	Utilities
WFC	Wells Fargo	Financials	EXC	Exelon	Utilities
ABBV	AbbVie	Health Care	SO	Southern Co.	Utilities

Note. The table reports the list of company tickers, company names, and industry sectors.

We estimate the integrated covariance for each trading day of 2009. On each of these days, we only consider the tickers that have at least 1,000 trades. Exploratory analysis (non reported in the paper) confirms the well-documented evidence of a CAPM-type factor structure in the panel. To this extent, our realized covariance estimation strategy consists of first decomposing the realized covariance in systematic and idiosyncratic covariance components and then regularizing the idiosyncratic part with the realized network. More precisely, we compute the realized covariance of the assets in the panel together with the SPY ticker (the proxy of the market), and then obtain the systematic and idiosyncratic components of the realized covariance of the assets on the basis of Formula 12. Finally, we apply the realized network regularization procedure to the idiosyncratic realized covariance. On each trading day of 2009, we estimate the realized network using three (idiosyncratic) realized covariance estimators: the RC, the TSRC, and the MRK.

6.2 | Realized network estimates

In this section we present the realized network estimation results. We first provide details for one specific date only that roughly corresponds to the middle of the sample (June 26, 2009) and then report summaries for all estimated networks in 2009. In the interest of space we report the TSRC estimator results only. The RC and MRK provide similar evidence.

We begin by showing in Figure 1 the heatmap of the idiosyncratic correlation matrix associated with the idiosyncratic realized covariance estimator on June 26. Note that the heatmap is constructed by sorting stocks by industry group and then by alphabetical order. The picture clearly shows that after netting out the influence of the market factor a fair amount of cross-sectional dependence is still present across stocks. Inspection of the heatmap reveals that the majority of estimated correlation coefficients are positive. The correlation matrix exhibits a block diagonal structure, hinting that correlation is stronger among firms in the same industry. On this date, the intra-industry group correlation is particularly strong for energy companies.

We estimate the realized network using the GLASSO and use the BIC to choose the optimal amount of shrinkage. The estimated network corresponding to the optimal λ has 244 edges, which correspond to approximately 5% of the total number of possible edges in the network on this date. The number of companies that have interconnections are 84 (roughly 90% of the total) and are all connected in a unique giant component.

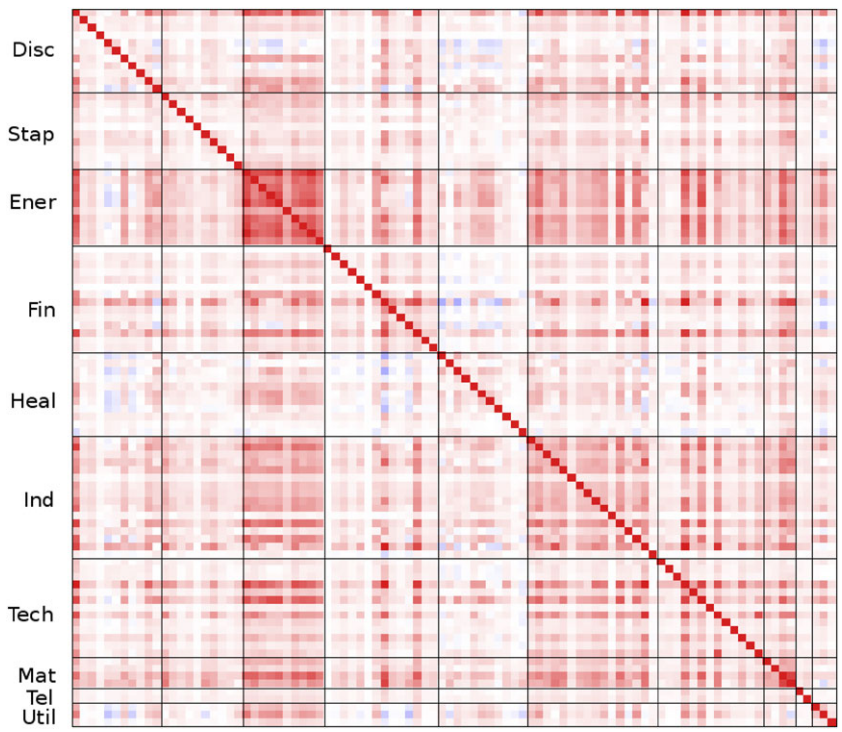

FIGURE 1 Idiosyncratic correlation heatmap on 2009-06-26. *Notes*: The figure shows the heatmap of the idiosyncratic realized correlation matrix on June 26, 2009, estimated using the TSRC estimator. Darker shades indicate higher correlations in absolute value [Colour figure can be viewed at wileyonlinelibrary.com]

It is useful to provide details on the amount of variability explained by the systematic and idiosyncratic components of the covariance matrix of the panel. To this extent, we introduce the systematic coefficient of determination, defined as

$$R^2_{F\,i} = \frac{\overline{\mathbf{B}}_i' \overline{\Sigma}_{FF} \overline{\mathbf{B}}_i}{\overline{\mathbf{B}}_i' \overline{\Sigma}_{FF} \overline{\mathbf{B}}_i + \hat{\sigma}_{I\,ii}},$$

which measures the amount of variability of asset i explained by the market factor. We also introduce the idiosyncratic coefficient of determination, defined as

$$R^2_{I\,i} = \frac{\hat{\sigma}_{I\,ii} - 1/\hat{k}_{I\,ii}}{\hat{\sigma}_{I\,ii}},$$

which measures the amount of variability of asset i explained by the remaining assets conditional on the market factor. On June 26, the average of the systematic R^2_F is equal to 27.1%, whereas the average of the idiosyncratic R^2_I (for those assets with at least one neighbor) is 12.5%. Overall, the systematic component is the most relevant one in terms of explained variability; however, the idiosyncratic component captures a nonnegligible portion of variability as well.

Figure 2 displays the idiosyncratic partial correlation network. A number of comments on the empirical characteristics of the network are in order. First, on this date, Google (GOOG) emerges as a particularly highly interconnected firm, with linkages spreading to several other industry groups. The estimated network also shows some degree of industry clustering; that is, linkages are more frequent among firms in the same industry group. In order to obtain better insights into the industry linkages in Table 3, we report the total number of links across industry groups. The table shows that firms in the energy and technology groups are particularly interconnected among each other. On the other hand, consumer discretionary, consumer staples and healthcare have few intra-industry linkages. In Figure 3 we report the degree distribution of the estimated network and the distribution of the nonzero partial correlations. As far as the degree distribution is concerned, the network exhibits the typical features of power law networks; that is, the number of connections

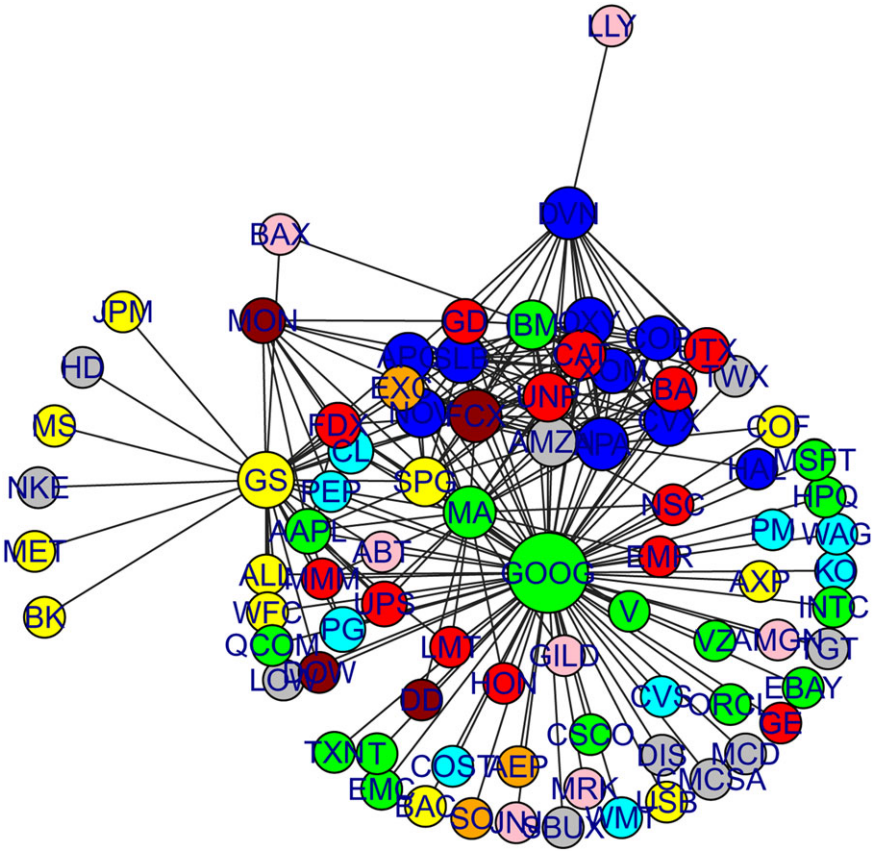

FIGURE 2 Idiosyncratic partial correlation network on 2009-06-26. *Notes*: The figure shows the optimal realized network estimated on June 26, 2009 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Links on 2009-06-26

	Disc	Stap	Ener	Fin	Heal	Ind	Tech	Mat	Tel	Util
Disc			7	4		6	11			1
Stap				3			13	3		
Ener	7		29	7	1	17	20	11		2
Fin	4	3	7	7	2	5	13	4		1
Heal			1	2			7			
Ind	6		17	5		4	31	6		
Tech	11	13	20	13	7	31	15	8	2	3
Mat		3	11	4		6	8	1		
Tel							2			
Util	1		2	1			3			

Note. The table reports the number of estimated links among industry groups on June 26, 2009.

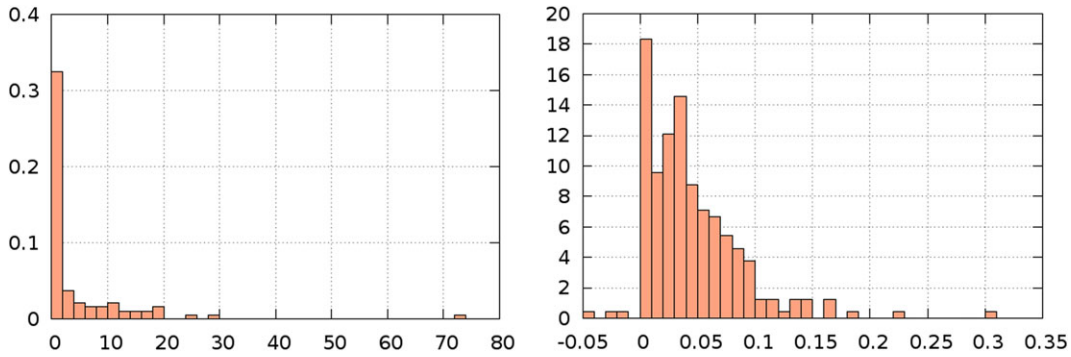

FIGURE 3 Degree and partial correlation distribution on 2009-06-26. *Notes*: The figure shows the degree distribution and the distribution of partial correlations on June 26, 2009 [Colour figure can be viewed at wileyonlinelibrary.com]

is heterogeneous and the most interconnected stocks have a large number of connections relative to the total number of links. The histogram of the partial correlation shows that the majority of the partial correlations are positive and that positive partial correlations are, on average, larger than the negative ones.

Last, we are interested in determining which companies are more interconnected and central in the network. We measure the degree of interconnectedness of a firm using different approaches: (i) the degree of a company in the network (i.e., the number of links); (ii) the sum of nonzero square partial correlations of a company; and (iii) the centrality index of the PageRank algorithm. The PageRank algorithm is a well-known network-based centrality index used by web search engines to rank web pages (see details in the Supporting Information Appendix). It turns out that the indices provide markedly close rankings. The rank correlations among the different measures are all above 0.9. We report the top 10 most central companies in Table 4 according to PageRank. The PageRank algorithm shows that Google is indeed the most central stock on this date.

We report a number of summary statistics for the sequence of networks estimated in 2009. First, in Figure 4 we report the proportion of links in the network throughout the year. The picture shows that sparsity is, on average, between 5% and 10% of the total possible number of linkages throughout the year. The plots show the sparsity rate vis-à-vis the VIX volatility index. The plot suggests that the network density is correlated to volatility: the higher the level of volatility, the more dense is the network.

Figure 5 shows the rolling monthly average of the total number of links of each industry group divided by the total number of possible edges. The plot omits the series for materials, telecom, and utilities due to their small size. Technology, energy, financial, and industrials are the most interconnected sectors throughout 2009, with technology being the most interconnected sector during the entire period. In order to provide more insights into the degree of concentration within each group, in Figure 6 we report the rolling monthly average of the concentration of links in each industry group

TABLE 4 Centrality on 2009-06-26

Rank	Ticker	Sector
1	GOOG	Information technology
2	FCX	Materials
3	APA	Energy
4	UNP	Industrials
5	IBM	Information technology
6	SLB	Energy
7	MA	Information technology
8	NOV	Energy
9	DVN	Energy
10	AMZN	Consumer discretionary

Note. The table reports the top tickers according to page rank on June 26, 2009.

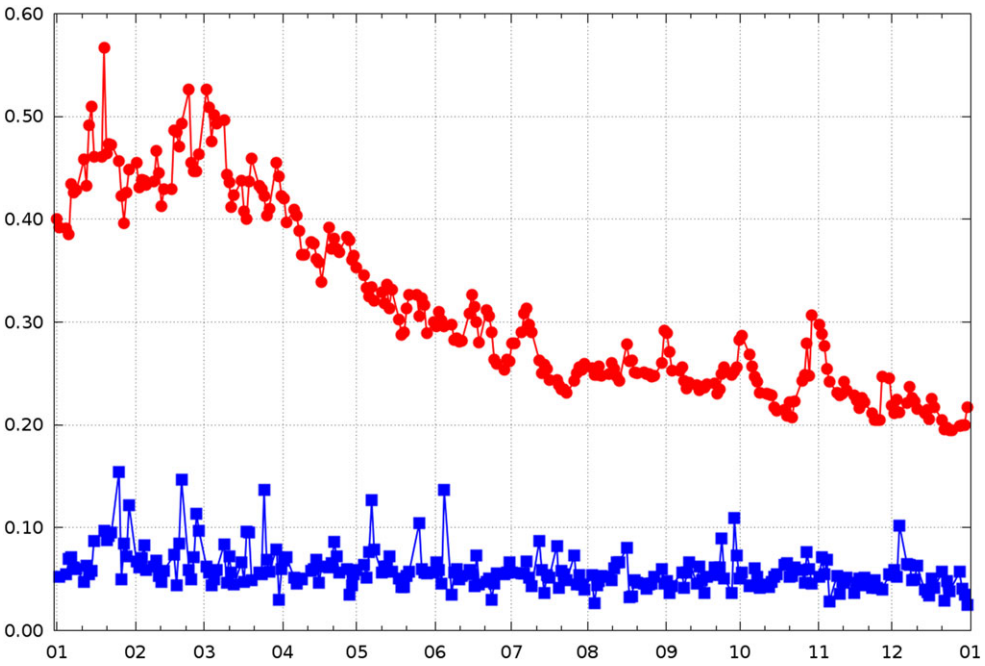

FIGURE 4 Sparsity versus volatility. *Notes*: The figure shows the sparsity of the estimated network (square) vis-à-vis the level of volatility measured by the VIX (circle) for each week of 2009 [Colour figure can be viewed at wileyonlinelibrary.com]

measured using the Herfindahl index.[5] Again, materials, telecom, and utilities are omitted from the graph. Technology and consumer discretionary are the most highly concentrated sectors. In particular, the consumer discretionary sector experiences a surge in the degree of concentration in the last part of the sample. In the case of the technology sector detailed inspection of the results reveals that this is driven by the fact that in 2009 Google is essentially the most interconnected ticker. For the consumer discretionary sector we have that Amazon progressively becomes more interconnected relative to the other companies in this sector throughout the sample. Industrials, on the other hand, have the smaller average concentration, in that the number of links is quite uniformly distributed across firms and no specific hub emerges among these tickers.

Overall results convey after conditioning on a one-factor structure that data still have a fair amount of cross-sectional dependence and that networks provide a useful device to synthesize such information. The main empirical features of the network are stable throughout 2009. Firms in the energy and industrials sectors are strongly interconnected. Technology companies, and Google in particular, are the most highly interconnected firms throughout the year.

[5]Let s_i denote the share of edges in the sub-graph of a given industry sector. Then the Herfindahl concentration index is defined as $H = \sum_i s_i^2$.

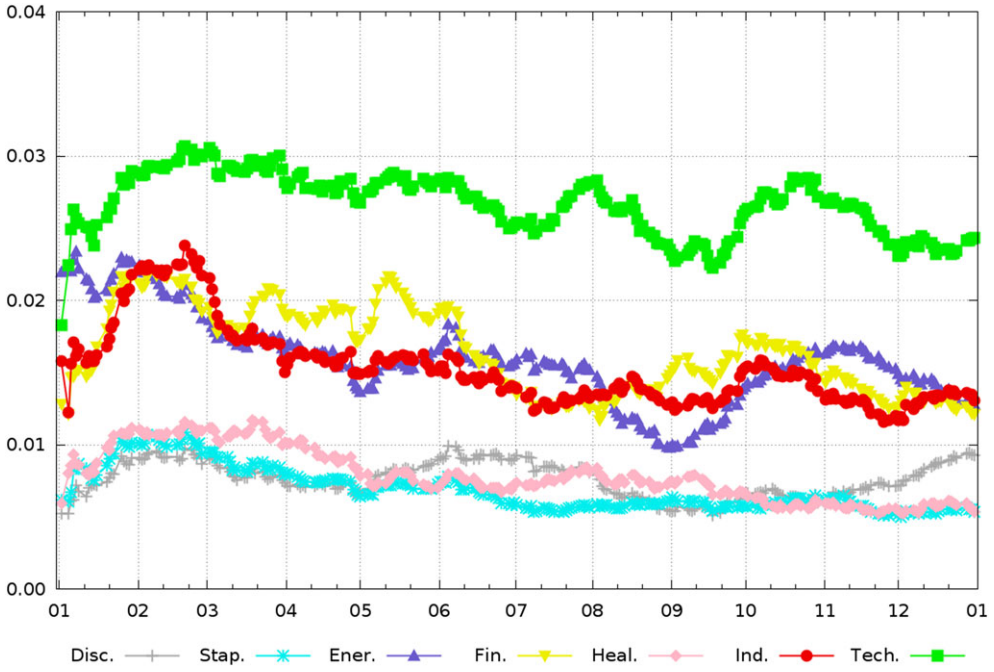

FIGURE 5 Sectorial links. *Notes*: The figure shows the number of linkages of the different industry groups over the total number of possible linkages for each week of 2009 [Colour figure can be viewed at wileyonlinelibrary.com]

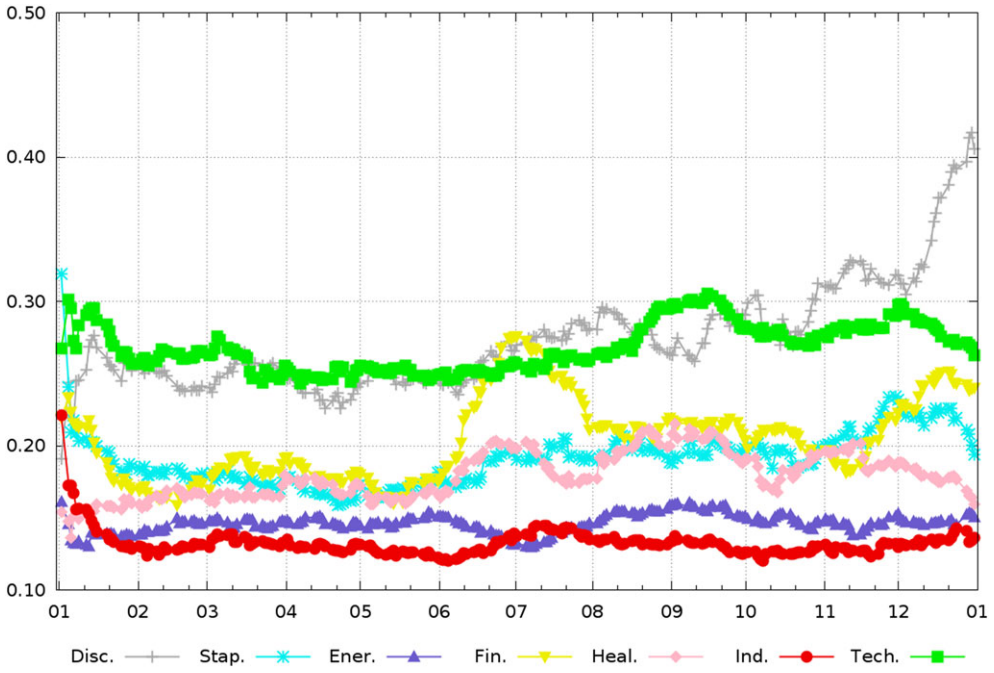

FIGURE 6 Sectorial concentration. Notes: The figure shows the link concentration (measured using the Herfindahl index) of the different industry for each week of 2009 [Colour figure can be viewed at wileyonlinelibrary.com]

6.3 | Predictive analysis

In order to assess the ability of the regularized network methodology to provide precise estimates of the integrated covariance we carry out an asset allocation prediction exercise (Engle & Colacito, 2006; Hautsch et al., 2012; 2015).

TABLE 5 GMV forecasting

	No regular.	Diagonal	Network	Factor	Shrinkage	Block-factor
RC	13.40	13.74	12.09	12.11	12.03	13.24
		−0.70	1.41	2.46**	2.65***	0.26
TSRC	13.05	13.52	11.96	12.34	12.16	12.29
		−0.74	1.88*	1.34	1.73*	0.88
MRK	12.95	13.89	11.65	13.03	10.46	11.15
		−1.82*	2.17**	−0.12	4.19***	2.30**

Note. The table reports the results of the GMV forecasting comparison exercise. The table reports the annualized out-of-sample volatilities of the GMV portfolios constructed for the unregularized RC, TSRC, and MRK estimators (No regular.) as well as their regularized versions (Diagonal, Network, Factor, Shrinkage, Block-factor), together with the test statistic of the Diebold–Mariano predictive ability test (asterisks denote significance at the standard significance levels).

The forecasting exercise is designed as follows. For each trading day of 2009, we construct the Markowitz global minimum variance (GMV) portfolio weights using the formula

$$\widehat{\mathbf{w}} = \frac{\widehat{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\Sigma}^{-1}\mathbf{1}}, \tag{15}$$

where $\mathbf{1}$ is an n-dimensional vector of ones and $\widehat{\Sigma}$ denotes some estimator of the integrated covariance over a given day. The resulting GMV portfolio weights are then used to construct a portfolio that is held until the following trading day. The exercise is carried out for each trading day from January 2, 2009 to December 31, 2009 (which corresponds to 252 days). The performance of different covariance estimators is evaluated by assessing which estimator delivers the smallest out-of-sample GMV portfolio variance. Note that we are implicitly relying on a random walk forecasting model to project the covariance of the assets and that prediction accuracy could be further enhanced by specifying an appropriate dynamic model.

The set of estimators we consider is based on the systematic/idiosyncratic decomposition of the covariance matrix:

$$\widehat{\Sigma} = \overline{\mathbf{B}}\overline{\Sigma}_{FF}\overline{\mathbf{B}}' + \widehat{\Sigma}_I, \tag{16}$$

and differ on the choice of estimator of the idiosyncratic realized covariance matrix $\widehat{\Sigma}_I$. The set of candidate idiosyncratic realized covariance estimators contains: (i) unregularized covariance estimator; (ii) constrained covariance estimator, obtained by setting all the off-diagonal elements of the unregularized covariance estimator to zero; (iii) realized network estimator; (iv) factor-regularized covariance estimator (see Equation 13) based on three factors; (v) shrinkage covariance estimator of Ledoit and Wolf (2004) (see Equation refeqn:lw); and (vi) block-factor regularized estimator, obtained by applying factor regularization of Equation 13 based on one factor to each industry block and setting the rest of the covariance matrix to zero. The exercise is carried out using the MRK, TSRC, and RC estimators.

We report summary results on the forecasting exercise in Table 5. The table shows the average annualized volatility of the GMV portfolios as well as the Diebold–Mariano test statistic. Asterisks denote significance of a Diebold–Mariano equal predictive ability test (Diebold & Mariano, 1995; Engle & Colacito, 2006) against the unregularized estimator. The three different covariance estimators deliver analogous inference. The constrained estimator that ignores cross–sectional dependence in the idiosyncratic realized covariance matrix performs worst than the baseline unconstrained estimator. Regularization improves over the unregularized estimator in the vast majority of cases. The shrinkage and realized network regularization schemes provide the best out–of–sample results. The factor and block-factor regularization also improve out-of-sample forecasting but the significance is weaker. We interpret this as the consequence that after controlling for the market factor there is only weak evidence of the presence of additional factors. The limited success of the block-factor regularization scheme may be due to the fact that, while some sectors exhibit strong dependence (industrials), a large number of stocks in other sectors do not (consumer discretionary, consumer staples, and healthcare).

7 | CONCLUSIONS

In this work we propose a regularization procedure for realized covariance estimators. The regularization consists of shrinking the off-diagonal elements of the inverse realized covariance matrix to zero using a LASSO-type penalty. Since

estimating a sparse inverse realized covariance matrix is equivalent to detecting the partial correlation network structure of the daily log-prices, we call our procedure realized network. The technique is specifically designed for the two-scales realized covariance (TSRC) and the multivariate realized kernel (MRK) estimators based on refresh time sampling, which are state-of-the-art consistent covariance estimators that allow for market microstructure effects and asynchronous trading. We establish the large-sample properties of the procedure estimator and show that the realized network consistently estimates the inverse integrated covariance matrix and consistently detects the nonzero partial correlations of the network. An empirical exercise is used to highlight the usefulness of the procedure, and an out-of-sample GMV portfolio asset allocation exercise is carried out to compare our procedure against a number of benchmarks. Results convey that realized network enhances the prediction properties of classic realized covariance estimators and performs favorably relative to a set of alternative regularization procedures.

ACKNOWLEDGEMENTS

We are grateful for useful comments to Peter Hansen, Jean-David Fermanian, Nour Meddahi, Kevin Sheppard, and participants at the "Measuring and Modeling Financial Risk with High Frequency Data" conference (Florence, June 19–20, 2014); "Barcelona GSE Summer Forum: Time Series Analysis in Macro and Finance" workshop (Barcelona, June 19–20, 2014); "7th Hedge Fund Research Conference" (Paris, January 22–23, 2015); "High Dimensional Time Series in Macroeconomics and Finance" conference at IHS (Vienna, May 21–22, 2015); "Barcelona GSE Summer Forum: High Frequency Financial Econometrics" workshop (Barcelona, June 11–12, 2015); "The Society for Financial Econometrics Eighth Conference" (Aarhus, June 23–25, 2015). Christian Brownlees acknowledges financial support from the Spanish Ministry of Science and Technology (Grant MTM2012-37195), Analysis of Big Data in Economics and Empirical Applications—2016 BBVA Foundation Grants for Scientific Research Teams, and from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2011-0075). Eulàlia Nualart's research is supported in part by the European Union programme FP7-PEOPLE-2012-CIG under grant agreement 333938.

REFERENCES

Acemoglu, D., Carvalho, V., Ozdaglar, A., & Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. *Econometrica*, *80*, 1977–2016.

Aït-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies*, *18*(2), 351–416.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*, 579–625.

Bandi, F. M., & Russell, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics*, *79*, 655–692.

Banerjee, O., & Ghaoui, L. E (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, *9*, 485–416.

Barigozzi, M., & Brownlees, C. (2013). NETS: Network estimation for time series. (*Technical report*). Barcelona, Spain: Barcelona GSE.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, *76*, 1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2009). Realised kernels in practice: Trades and quotes. *Econometrics Journal*, *12*, 1–32.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2011). Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, *162*(2), 149–169.

Barndorff-Nielsen, O. E., & Shephard, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, *72*(3), 885–925.

Brownlees, C. T., & Gallo, G. M. (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics and Data Analysis*, *51*, 2232–2245.

Corsi, F., Peluso, S., & Audrino, F. (2014). Missing in asynchronicity: A Kalman-em approach for multivariate realized covariance estimation. *Journal of Applied Econometrics*, *30*, 377–397.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157–175.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–263.

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

Engle, R., & Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business and Economic Statistics*, *24*, 238–253.

Fan, J., Fan, Y., & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, *147*, 186–197.

Fan, J., Li, Y., & Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, *107*(497), 412–428.

Fan, J., Liao, Y., & Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, *39*, 3320–3356.

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B*, *75*, 603–680.

Friedman, J., Hastle, T., Hofling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*, 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2011). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441.

Hautsch, N., Kyj, L. M., & Malec, P. (2015). Do high-frequency data improve high-dimensional portfolio allocations? *Journal of Applied Econometrics*, *30*, 263–290.

Hautsch, N., Kyj, L. M., & Oomen, R. C. A. (2012). A blocking and regularization approach to high dimensional realized covariance estimation. *Journal of Applied Econometrics*, *27*(4), 625–645.

Hautsch, N., Schaumburg, J., & Schienle, M. (2014a). Financial network systemic risk contributions. *Review of Finance*, *19*, 685–738.

Hautsch, N., Schaumburg, J., & Schienle, M. (2014b). Forecasting systemic impact in financial networks. *International Journal of Forecasting*, *30*(3), 781–794.

Kim, D., Wang, Y., & Zou, J. (2016). Asymptotic theory for large volatility matrix estimation based on high-frequency financial data. *Stochastic Processes and their Applications*, *126*(11), 3527–3577.

Laurent, S., Rombouts, J. V., & Violante, F. (2011). On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics*, *27*(6), 934–955.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*, 365–411.

Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, *40*, 1024–1060.

Lunde, A., Shephard, N., & Sheppard, K. (2016). Econometric analysis of vast covariance matrices using composite realized kernels and their application to portfolio choice. *Journal of Business & Economic Statistics*, *34*(4), 504–518.

Martens, M. (2004). Estimating unbiased and precise realized covariances. (Working paper), Econometric Institute, Erasmus University Rotterdam, Rotterdam, Netherlands.

Meinshausen, N., & Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, *34*, 1436–1462.

Peng, J., Wang, P., Zhou, N., & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, *104*, 735–746.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. Hoboken, NJ: Wiley.

Ravikumar, P., Wainwright, M. J., Raskutti, G., & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, *5*, 935–980.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, CA, pp. 157–164.

Tao, M., Wang, Y., & Zhou, H. H. (2013). Optimal sparse volatility matrix estimation for high-dimensional Itô processes with measurement errors. *Annals of Statistics*, *41*, 1816–1864.

Varneskov, R. T. (2016). Flat-top realized kernel estimation of quadratic covariation with nonsynchronous and noisy asset prices. *Journal of Business and Economic Statistics*, *34*(1), 1–22.

Wang, Y., & Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *Annals of Statistics*, *38*, 943–978.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19–35.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, *160*(1), 33–47.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Brownlees C, Nualart E, Sun Y. Realized networks. *J Appl Econ*. 2018;1–21. https://doi.org/10.1002/jae.2642