Augment Large Covariance Matrix Estimation With Auxiliary Network Information

Shuyi Ge[*], Shaoran Li[†], Oliver Linton[‡], and Weiguang Liu[§]

Faculty of Economics, University of Cambridge

June 14, 2021

Abstract

This paper aims to incorporate auxiliary information about the location of significant correlations into the estimation of high-dimensional covariance matrices. With the development of machine learning techniques such as textual analysis, granular linkage information among firms that used to be notoriously hard to get are now becoming available to researchers. Our Network Guided Estimator combines the banding and thresholding procedures with the help of augment information from other sources. Simulation results show that the new method has smaller estimation errors comparing with other methods in the literature. We empirically apply the Network Guided Estimator to estimate the covariance of the excess returns of SP500 stocks. The constructed global minimum variance portfolio has the smallest volatility among all competing methods.

1 Introduction and Literature Review

Covariance matrix estimation is an important area of research in both finance and statistics. Suppose we have independent observations $X_t = (X_{1t}, \dots, X_{Nt})^\intercal, t = 1, \dots, T$

[*]Author email:sg751@cam.ac.uk
[†]Author email:sl736@cam.ac.uk
[‡]Author email:obl20@cam.ac.uk
[§]Author email:wl342@cam.ac.uk

of a $N-$dimensional random vector $\mathbf{X}$ that has mean μ and variance $\Sigma_X = E((\mathbf{X}-\mu)(\mathbf{X}-\mu)^\intercal)$. The most straightforward estimator is the sample covariance estimator, which is defined as follows:

$$\hat{\Sigma}_X = \frac{1}{T}(X-\bar{X})(X-\bar{X})^\intercal = [\hat{\sigma}_{ij}]_{N\times N}, \tag{1}$$

where X is the $N \times T$ matrix of observations, $\bar{X} = \frac{1}{T}X1_T1_T^\intercal$ is the sample time series average, with 1_T being a $T\times 1$ vector of 1. However, in the high-dimensional case, where the dimension N is not negligible comparing to sample size T, the sample covariance matrix is ill-conditioned and inconsistent. Some structures need to be imposed on Σ_X, and regularization techniques need to be applied to make sure the estimator is reliable.

One of the structures that is often imposed in the high-dimensional settings is sparsity, which assumes that Σ_X is sparse (i.e., has lots of zeros or small elements) or conditionally sparse (i.e., has lots of zeros or small elements once we condition on some variables like the common risk factors). Given the sparsity structure, several estimation strategies have been proposed in the literature such as banding, tapering, shrinkage, thresholding, etc.

Banding and tapering are applicable when we can find a "distance" $d(i,j)$ between X_i and X_j. For example, in time series, a natural distance is $|i-j|$ and we have reasons to believe that a large distance between X_i and X_j will imply a lower correlation. Such structure is appropriate for applications where there are natural orderings of variables, such as time series, climatology and spectroscopy. Banding methods (Bickel and Levina, 2008b) keep only elements within a k-neighbourhood of each individual X_i, that is

$$\hat{\Sigma}_{B,k} = [\hat{\sigma}_{ij}\mathbf{1}_{d(i,j)\le k}],$$

whereas the tapering estimator does an element-wise multiplication of $\hat{\Sigma}_X$ with a tapering positive definite matrix $\mathbf{T}$ that has smaller element $\mathbf{T}_{ij}$ decreases with $d(i,j)$.

When we do not have the information about the distance between i and j, one of the viable approaches is thresholding. Let $s_\lambda(\cdot)$ be a generalized thresholding operator with thresholding level λ, such that for all $z \in \mathbb{R}$, (1). $|s_\lambda(z)| \le |z|$ for all z (2). $s_\lambda(z) = 0$ for $|z| \le \lambda$ (3). $|s_\lambda(z) - z| \le \lambda$, which incorporates commonly used thresholding operators such as hard thresholding, soft thresholding, SCAD, etc. Then a thresholding estimator is

$$\hat{\Sigma}_\lambda = [\tilde{\sigma}_{ij}] \quad \text{and} \quad \tilde{\sigma}_{ij} = \begin{cases} \hat{\sigma}_{ii} & i = j \\ s_\lambda(\hat{\sigma}_{ij}) & i \neq j \end{cases}$$

Bickel and Levina, 2008a develop the theory for universal thresholding, which assumes that the diagonal of Σ is uniformly bounded. Cai and W. Liu, 2011 propose an adaptive thresholding estimator. They relax the uniform boundedness assumption and account for the variance of the estimator of each $\hat{\sigma}_{ij}$ to establish entry-adaptive threshold level $\hat{\Sigma}_X$. Fan, Liao, and Mincheva, 2013 argue that common factors should be extracted first before applying thresholding when there are "extremely spiked" eigenvalues in $\hat{\Sigma}_X$ and the covariance matrix is conditionally sparse. Shu and Nan, 2019 obtain the convergence rate allowing for temporal dependence. Bickel and Levina, 2008a also compare the convergence rates of banding estimator and thresholding estimator. By utilizing the location information, banding estimator shows a superior convergence rate. See Fan, Liao, and H. Liu, 2015 for a review on estimation of large dimensional covariance matrices.

Apart from element-wise regularization methods, Ledoit and Wolf, 2004a and Ledoit and Wolf, 2012 have proposed linear and nonlinear shrinkage estimators that apply shrinkage to the eigenvalues of the sample covariance matrix. The linear shrinkage does that by finding the linear combination of sample covariance and a well-conditioned matrix such as the identity matrix and nonlinear shrinkage estimator corrects the eigenvalues using the asymptotic Marcenko–Pastur distribution. Shrinkage estimators have been successfully applied in portfolio construction (Ledoit and Wolf, 2004b, Ledoit and Wolf, 2017).

Although in general we won't have an ordering or the distance between i and j, we do have some idea about who might be connected with whom using auxiliary information apart from the observations of X. To proxy for pairwise connectivity among entities, several methods have been proposed. Hoberg and Phillips, 2016 use textual analysis to identify peers. Kaustia and Rantala, 2013 identify peers by analyst co-coverage, and Ge and O. B. Linton, 2021 identify peers using business news co-mentioning. The network information gathered from these sources can help us to identify the locations of non-zero elements. Similar location-based thresholding ideas have applied in Fan, Furger, and Xiu, 2016 and Brownlees, Gumundsson, and Lugosi, 2020. Fan, Furger, and

Xiu, 2016 apply a hard thresholding method in a way that $\sigma_{ij} = 0$ when i and j are from different sector/industry. The network they use is a time-invariant block-diagonal matrix, and our method can accommodate more general and flexible network information. Brownlees, Gumundsson, and Lugosi, 2020 first detect community structure using a spectral clustering-based procedure, and then apply a block-by-block thresholding to the off-diagonal elements of $\hat{\Sigma}_X$. In particular, they do not apply thresholding to $\hat{\sigma}_{ij}$ if i and j are from the same community.

In this paper, we use granular network information gathered from other sources. We argue that we can incorporate such auxiliary information in the estimation of the covariance matrix Σ when they help reveal the locations of the large elements (or nonzero elements in the strictly sparse case). We propose a *Network Guided Estimator* that combines the banding and thresholding procedures. Given a network $\hat{L}$, we keep the elements $\hat{\sigma}_{ij}$ where $\hat{L}_{ij} = 1$ and apply generalized thresholding operator on the elements $\hat{\sigma}_{ij}$ where $\hat{L}_{ij} = 0$. This estimator will relax the sparsity conditions on the covariance Σ. In the simulations, we show that the Network Guided Estimator has smaller estimation error comparing to the sample covariance, linear shrinkage, nonlinear shrinakge and the universal thresholding estimators.

As an empirical application, we use the Network Guided Estimator to estimate the covariance of excess returns of SP500 stocks, and construct the global minimum variance portfolio. The first source of location information comes from the new-implied network. It has been documented that common news coverages reveal information about linkages among companies, which are related to many economically important relationships like business alliances, partnerships, banking and financing, customer-supplier, and production similarity (Scherbina and Schlusche, 2015, Schwenkler and Zheng, 2019). Ge and O. B. Linton, 2021 document that stocks linked by news co-mentioning exhibit additional co-movement beyond what can be explained by common risk factors. Same as Ge and O. B. Linton, 2021, we use news data from RavenPack Equity files Dow Jones Edition for the period from the beginning of 2004 to the end of 2015. This comprehensive news dataset combines relevant contents from multiple sources, including Dow Jones Newswires, Wall Street Journal, and Barron's Market-Watch, which produce the most actively monitored streams of news articles in the financial system. We identify linkages among firms by news co-mentioning.

As a simple starting point, in each rolling window, we set $\hat{L}_{ij} = 1$ if the firms (i, j)

are co-mentioned more than 50 times during the past year and their current sample correlation reaches a certain level. We apply the Network Guided Estimator to the de-factored excess returns and show that the global minimum variance portfolio that we constructed have smaller standard deviation than portfolios constructed using other methods. This empirical analysis is still preliminary and we are looking into ways to construct better estimator $\hat{L}_{ij}$.

2 Estimator and Convergence Rate

Assume we have observations $X = (X_1, \ldots, X_T)$, where X_t are independent drawn from a N-dimensional distribution F with mean μ and variance Σ. In Bickel and Levina, 2008a, they consider the following uniformity class of covariance matrices:

$$\mathcal{U}_\tau(q, c_0(p), M) = \left\{ \Sigma : \sigma_{i,i} \leq M, \sum_{j=1}^{p} \left|\sigma_{ij}\right|^q \leq c_0(p), \text{ for all } i \right\}$$

And the convergence rate will depend on $c_0(p)$ and q. Notice that in order to bound the i-th row sparsity index $\mathcal{S}_i(q) := \sum_{j=1}^{p} \left|\sigma_{ij}\right|^q$ by $c_0(p)$, the coefficient q is important. Suppose $q = 0$, then $\mathcal{S}_i = \#\left\{\sigma_{ij} \neq 0\right\}$. If the i-th row $\Sigma_{i\cdot}$ has a lot of nonzero but small elements, then to bound $\mathcal{S}_i(0)$ would require a higher $c_0(p)$. On the other hand when $q \longrightarrow 1$, the large elements of $\Sigma_{i\cdot}$ will dominate. Hence if Σ is sparse in the sense that it contains a small number of relatively large elements and a large number of small elements, it's advisable to state conditions separately for these elements, and we consider the following uniformity class:

$$\mathcal{U}(q, c_0, c_1, M, L) = \left\{ \Sigma : \sigma_{ii} \leq M, \sum_j L_{ij}^1 \leq c_1(p), \sum_j L_{ij}^0 \left|\sigma_{ij}\right|^q \leq c_0(p) \text{ for all } i \right\}$$

where L_{ij} represents the location of the large elements and for $s \in \{0, 1\}$, $L_{ij}^s = \mathbf{1}_{L_{ij}=s}$ This uniformity class controls the number of the large elements at locations $((i, j) : L_{ij}^1 = 1)$ and the growth rate of the remaining small elements.

Of course, a priori we don't know the location of the large elements, but suppose we have observations from auxiliary dataset that allow us to form an estimator $\hat{L}$ for L, independent of the sample X, we can design an estimator that takes into account the addition information in $\hat{L}$. A simple choice is to do banding based on the location information and apply thresholding on the reminder terms. Here we define a *Network Guided Estimator* to be

$$T_{L,\lambda}(\hat{\Sigma}) = \left[s_{L,\lambda}(\hat{\sigma}_{ij})\right]_{N\times N}$$
$$s_{L,\lambda}\left(\sigma_{ij}\right) = \begin{cases} \sigma_{ij} & \text{if} \quad i = j \quad \text{or} \quad L_{ij} = 1 \\ s_\lambda(\sigma_{ij}) & \text{otherwise} \end{cases}$$

where $s_\lambda(x)$ is the generalized thresholding operator and $\hat{\sigma}_{ij}$ are elements of the sample covariance matrix, then the feasible Network Guided Estimator is $T_{\hat{L},\lambda}\left(\hat{\Sigma}\right)$

Assumption 1. We make the following assumptions:

1. $\max_{ij}\left|\hat{\sigma}_{ij} - \sigma_{ij}\right| = O_p(\sqrt{\log N/T})$;
2. $\max_{ij}\left|\hat{L}_{ij} - L_{ij}\right| = O_p(k_T)$ where $k_T \to 0$ as $T \to \infty$.

Remark. 1. The first assumption can be verified in various settings, for example, if F is Gaussian or sub-Gaussianv(Cai and W. Liu, 2011). We can even replace the independence assumption and allow for temporal dependence (see Lemma A.2 in Shu and Nan, 2019).

2. The second assumption appears restrictive, but since L_{ij} are estimated independently from different datasets, we believe it's not too strigent to require that for each (i, j), L_{ij} can be estimated consistently. In addition, simulation shows that as long as we don't make too much II-type error: $\hat{L}_{ij} = 1$ when $L_{ij} = 0$, using the additional information still improves the performance.

   Given that the estimation of $\hat{L}_{ij}$ is independent of the sample (X_t), then perhaps we can find a less restrictive condition for the result to hold.

Theorem 1. *Suppose F is Gaussian and for sufficiently large M,*

$$\lambda = M\sqrt{\frac{\log N}{T}}$$

and $\frac{\log N}{T} \to 0$ *as* $T \to \infty$*, then for the operator norm* $\|M\| = \max_j \left|\lambda_j(M)\right|$ *where* $\lambda_1, \ldots, \lambda_N$ *are the eigenvalues of* M*, we have*

$$\left\|T_{\hat{L},\lambda}\left(\hat{\Sigma}\right) - \Sigma\right\| = O_p\left(c_1(p)\sqrt{\frac{\log N}{T}} + c_0(p)\left(\frac{\log N}{T}\right)^{\frac{1-q}{2}}\right).$$

3 Simulations

We demonstrate the Network Guided Estimator and examine its small-sample performance using the following simulations. First, we consider the case where the true covariance Σ comes from an AR(1) model. So for $\{(i,j) : i = 1, \ldots, N, j = 1, \ldots, N\}$, $\sigma_{ij}^2 = \sigma_i \sigma_j \rho_{ij}$ and $\rho_{ij} = \rho^{|i-j|}$, we take $N = 200$ and

$$\sigma_{ij} = 3 * \rho^{|i-j|}.$$

Assume we observe a matrix $\hat{L}$ indicating the location of highly correlated pairs $L_{ij} = \mathbf{1}\left\{\rho_{ij} \geq l\right\}$. Conditional on $L_{ij} = 1$, we observe $\hat{L}_{ij} = 1$ with probability p and conditional on $L_{ij} = 0$, $\hat{L}_{ij} = 1$ with probability q. Hence p, q reflect the probability of missing important locations (type II error) and including falsely important locations (type I error) respectively.

We then generate $T = 100$ independent draws of observations X_t from $N(0, \Sigma)$ and estimate Σ using 1. Sample covariance; 2. Linear Shrinkage estimator; 3. Nonlinear Shrinkage estimator; 4. Universal thresholding on the correlation; 5. and Network Guided Estimator. We now compare their performance. It's worth collecting here the parameters that we will adjust in the experiments in Table 1

In Table 3, 4 and 5, we show the estimation errors of these estimators in terms of the operator norm, the Frobenius norm and the matrix 1-norm, when we simulate using different ρ and thresholding level λ and fix the other parameters at $l = 0.3, p = 1$ and $q = 0$. Here we have taken the thresholding operator to be soft thresholding. It can be seen that for all these norms, when the true covariance matrix is not too dense, Network Guided Estimator outperforms all the competitors given a good choice of λ. When the true covariance matrix is more sparse, indicated by smaller ρ, thresholding methods become more appealing and when the true covariance matrix is denser, the sample covariance estiamtor shows better performance compared with other bench-

Parameter	Description
ρ	Determines how strong the correlation is and the sparsity of the covariance matrix Σ
l	Observation level, determines how we classify a pair (i, j) as important, i.e., $L_{ij} = \mathbf{1}\{\rho_{ij} > l\}$.
p	Conditional on $L_{ij} = 1$, the probability of actually observing $\hat{L}_{ij} = 1$.
q	Conditional on $L_{ij} = 0$, the probability of observing $\hat{L}_{ij} = 1$
λ	The Threshold level when we apply generalized thresholding operator on σ_{ij} where $\hat{L}_{ij} = 0$.

Table 1: Description of varying parameters.

mark models. Thanks to the accurate location information, Network Guided Estimator is able to balance these two estimators.

Another tuning parameter in the Network Guided Estimator is the choice of l, which can determine whether a link should be reserved regardless of the information from the statistic. Then we consider simulations with varying observation levels l. In Figure 1, 2, 3, when we set observation level equal to 0, the Network Guided Estimator will be the same as the sample covariance estimator, on the other extreme, when observation level is set to 1, the Network Guided Estimator is equivalent to universal thresholding. In between these cases, when we have information about the locations of the important pairs, we have a range where the estimation errors can be lowered.

Then we show the effects of errors in estimating the L_{ij} by varying the parameters p and q. In Table 6, 7, 8, we have when $p = q = 0$ the estimation error of the universal thresholding estimator, and $p = q = 1$ the sample covaraince estimation error. As we can see, as long as q is not too large, the estimation error will be smaller when we have a higher probability p of observing the true large elements. It should be noted that q in fact cannot be very large, given that the whole matrix is sparse.

4 Empirical Studies

4.1 Global Minimum Variance Portfolio

We apply the Network Guided Estimator to a portfolio management similar to Ledoit and Wolf, 2004b. We collect daily return data on SP500 stocks and Fama-French 3 factors and the risk-free rates from 2004 to 2015 from Center for Research in Security Prices (CRSP).

Assume that the excess returns follow the following factor model

$$X_{it} = B_i' F_t + \varepsilon_{it},$$

and we assume that $\Sigma = [E\varepsilon_i\varepsilon_j]_{1\leq i,j\leq N}$ is sparse.

We do a rolling-window analysis: each window consists of an estimation period of 252 days and a testing period of 21 days. In the estimation period, we estimate the factor loadings by linear time series regression of excess return X_{it} on the factors F_t, hence allowing the betas B_i to vary over time, and we find the de-factored excess return by

$$\hat{\varepsilon}_{it} = X_{it} - \hat{B}_i' F_t$$

and in order to estimate the covariance matrix of $X = (X_1, \ldots, X_N)'$, we have, under the assumption that ε's are independent of F_t,

$$\Sigma_X = B\Sigma_F B' + \Sigma_\varepsilon$$

We replace the factor covariance component by $\hat{B}\hat{\Sigma}_F\hat{B}$, where $\hat{\Sigma}_F$ is the sample covariance of factors in that period, and we estimate Σ_ε by the Network Guided Estimator applied to $\hat{\Sigma}_\varepsilon = \frac{1}{T}\sum_t \hat{\varepsilon}_t\hat{\varepsilon}_t'$.

In order to apply the Network Guided Estimator, we consider a preliminary estimator $\hat{L}_{ij}$ using the RavenPack news data, where $\hat{L}_{ij} = 1$ if the pair of firms (i, j) have been co-mentioned more than 50 times during the past year and the current sample correlation is higher than 0.5 to keep G sparse and mitigate the noisy observations.

We select the thresholding parameter using cross-validation with the constraint that the resulting estimate is positvie definite. When the thresholding level becomes higher, the resulting estimate becomes more sparse, in the limit, it'll be a diagonl matrix

Table 2: The mean, standard deviation, minimum, maximum and Sharpe ratio of the holding period return of the global minimum variance portfolio constructed using different covariance estimators.

	mean	std	min	max	Sharpe Ratio
New Method	1.0080	0.0255	0.9175	1.0834	0.3796
Linear Shrinkage	1.0044	0.0264	0.8752	1.0630	0.3792
Universal Thresholding	1.0058	0.0263	0.8836	1.0750	0.3793
Equally Weighted	1.0117	0.0667	0.8073	1.3069	0.3803

and positive definite, see discussion in Fan, Liao, and H. Liu, 2015.

After using the data from estimation period to estimate $\hat{B}, \hat{\Sigma}_F, \hat{\Sigma}_\varepsilon$, we construct

$$\hat{\Sigma}_X = \hat{B}\hat{\Sigma}_F\hat{B}' + \hat{\Sigma}_\varepsilon$$

and then construct the *global minimum variance* portfolio with weights given by

$$w = \frac{\hat{\Sigma}_X \mathbf{1}}{\mathbf{1}'\hat{\Sigma}_X \mathbf{1}}$$

where $\mathbf{1}$ is a conforming vector of ones. We collect the portfolio return over the next 21-day testing period. This conludes one of the rolling windows. Then we move forward 21 days and repeat this exercise. Using 2004-2015 daily data, we can construct a daily portfolio return from 2005 to 2015, where the portfolio is rebalanced every 21 days. We compute the holding-period return of this portfolio and show in Table 2 the standard deviation together with mean and the Sharpe ratio. We compare it with global minimum variance portfolio constructed using linear shrinkage and universal thresholding and the equally weighted portfolio. It's worth mentioning that given we are comparing global minimum variance portfolio, the standard deviation is the most relevant indicator of performance.

5 Conclusion

This paper considers the problem of incorporating auxialiary data such as textual-analysis data into the estimation of large covariance matrices. It can be shown that by incorporating information about locations of important links we can relax the sparsity conditions of the thresholding estimators and in simulations we show the proposed Network Guided Estimator has superior performance in finite samples. We have also applied the Network Guided Estimator in the construction of minimum-variance portfolio in a preliminary empirical application. There are several improvements that we are undertaking.

Firstly, the construction of good estimator $\hat{L}_{ij}$ for the important locations is an important question. It's apparent from the simulations that the quality of $\hat{L}_{ij}$ will affect the estimation error. We have used a straightforward estimator in the empirical study, but it's not completely satisfactory and we need a more systematic way of constructing the $\hat{L}$. Secondly, we are expanding the set of auxialiary networks beyond the RavenPack news data to include Hoberg's similarity score and IBES co-covarage data, as well as applying the technique on a larger dataset.

Although here we use network information to the estimation of large static covariance matrix, similar ideas can be extended to the estimation of large dynamic covariance matrix. For example, dynamic network information could be incorporated into the conditioning information set like in Chen, Li, and O. Linton, 2019.

Appendix

5.1 Tables and Figures

Table 3: The estimation error of various estimators in terms of the Operator Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
0.70	0.0	20.48	11.43	11.16	20.48	20.48
	0.1	20.48	11.43	11.16	16.72	17.01
	0.2	20.48	11.43	11.16	13.28	13.86

Continued on next page

Table 3: The estimation error of various estimators in terms of the Operator Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	0.3	20.48	11.43	11.16	10.30	11.15
	0.4	20.48	11.43	11.16	8.43	8.97
	0.5	20.48	11.43	11.16	8.41	7.36
	0.6	20.48	11.43	11.16	8.60	6.69
	0.7	20.48	11.43	11.16	8.92	6.67
	0.8	20.48	11.43	11.16	9.28	6.70
	0.9	20.48	11.43	11.16	9.63	6.73
	1.0	20.48	11.43	11.16	10.00	6.79
0.80	0.0	18.82	17.07	16.17	18.82	18.82
	0.1	18.82	17.07	16.17	14.82	15.32
	0.2	18.82	17.07	16.17	13.46	12.27
	0.3	18.82	17.07	16.17	13.00	11.21
	0.4	18.82	17.07	16.17	12.90	10.48
	0.5	18.82	17.07	16.17	13.07	10.04
	0.6	18.82	17.07	16.17	13.47	9.82
	0.7	18.82	17.07	16.17	14.03	9.78
	0.8	18.82	17.07	16.17	14.67	9.83
	0.9	18.82	17.07	16.17	15.38	9.96
	1.0	18.82	17.07	16.17	16.09	10.10
0.90	0.0	42.51	27.24	28.43	42.51	42.51
	0.1	42.51	27.24	28.43	35.92	37.50
	0.2	42.51	27.24	28.43	29.98	33.18
	0.3	42.51	27.24	28.43	24.91	29.74
	0.4	42.51	27.24	28.43	20.75	27.19
	0.5	42.51	27.24	28.43	18.41	25.40

Continued on next page

Table 3: The estimation error of various estimators in terms of the Operator Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	0.6	42.51	27.24	28.43	20.55	24.18
	0.7	42.51	27.24	28.43	22.82	23.34
	0.8	42.51	27.24	28.43	25.09	22.75
	0.9	42.51	27.24	28.43	27.26	22.29
	1.0	42.51	27.24	28.43	29.30	21.91
0.95	0.0	44.59	41.02	38.02	44.59	44.59
	0.1	44.59	41.02	38.02	38.02	40.90
	0.2	44.59	41.02	38.02	37.36	37.59
	0.3	44.59	41.02	38.02	39.13	34.71
	0.4	44.59	41.02	38.02	41.53	32.25
	0.5	44.59	41.02	38.02	44.41	31.65
	0.6	44.59	41.02	38.02	47.61	32.02
	0.7	44.59	41.02	38.02	50.96	32.55
	0.8	44.59	41.02	38.02	54.39	33.16
	0.9	44.59	41.02	38.02	57.74	33.75
	1.0	44.59	41.02	38.02	60.95	34.27
0.99	0.0	29.02	35.94	28.34	29.02	29.02
	0.1	29.02	35.94	28.34	34.93	29.06
	0.2	29.02	35.94	28.34	44.67	29.68
	0.3	29.02	35.94	28.34	55.51	30.89
	0.4	29.02	35.94	28.34	66.94	32.48
	0.5	29.02	35.94	28.34	78.72	34.28
	0.6	29.02	35.94	28.34	90.71	36.17
	0.7	29.02	35.94	28.34	102.79	38.01
	0.8	29.02	35.94	28.34	114.89	39.76

Continued on next page

Table 3: The estimation error of various estimators in terms of the Operator Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	0.9	29.02	35.94	28.34	126.93	41.37
	1.0	29.02	35.94	28.34	138.84	42.75

Table 4: The estimation error of various estimators in terms of the Frobenius Norm

ρ	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
0.70	0.0	59.43	42.23	41.61	59.43	59.43
	0.1	59.43	42.23	41.61	49.91	49.85
	0.2	59.43	42.23	41.61	41.94	41.67
	0.3	59.43	42.23	41.61	35.66	34.97
	0.4	59.43	42.23	41.61	31.20	29.81
	0.5	59.43	42.23	41.61	28.56	26.18
	0.6	59.43	42.23	41.61	27.54	23.93
	0.7	59.43	42.23	41.61	27.78	22.81
	0.8	59.43	42.23	41.61	28.83	22.48
	0.9	59.43	42.23	41.61	30.33	22.59
	1.0	59.43	42.23	41.61	32.08	22.98
0.80	0.0	62.54	47.59	46.59	62.54	62.54
	0.1	62.54	47.59	46.59	52.60	52.80
	0.2	62.54	47.59	46.59	44.30	44.52
	0.3	62.54	47.59	46.59	37.89	37.85
	0.4	62.54	47.59	46.59	33.55	32.85

Continued on next page

Table 4: The estimation error of various estimators in terms of the Frobenius Norm

ρ	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	0.5	62.54	47.59	46.59	31.32	29.48
	0.6	62.54	47.59	46.59	30.95	27.58
	0.7	62.54	47.59	46.59	31.93	26.78
	0.8	62.54	47.59	46.59	33.75	26.71
	0.9	62.54	47.59	46.59	36.02	27.05
	1.0	62.54	47.59	46.59	38.49	27.59
0.90	0.0	63.06	53.18	52.83	63.06	63.06
	0.1	63.06	53.18	52.83	53.98	54.60
	0.2	63.06	53.18	52.83	46.91	47.84
	0.3	63.06	53.18	52.83	42.16	42.89
	0.4	63.06	53.18	52.83	39.78	39.67
	0.5	63.06	53.18	52.83	39.57	37.96
	0.6	63.06	53.18	52.83	41.08	37.43
	0.7	63.06	53.18	52.83	43.74	37.69
	0.8	63.06	53.18	52.83	47.08	38.41
	0.9	63.06	53.18	52.83	50.74	39.33
	1.0	63.06	53.18	52.83	54.57	40.32
0.95	0.0	57.97	52.58	51.77	57.97	57.97
	0.1	57.97	52.58	51.77	51.42	52.21
	0.2	57.97	52.58	51.77	47.65	48.39
	0.3	57.97	52.58	51.77	46.74	46.40
	0.4	57.97	52.58	51.77	48.35	45.96
	0.5	57.97	52.58	51.77	51.82	46.66
	0.6	57.97	52.58	51.77	56.46	48.04
	0.7	57.97	52.58	51.77	61.73	49.73

Continued on next page

Table 4: The estimation error of various estimators in terms of the Frobenius Norm

ρ	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	0.8	57.97	52.58	51.77	67.33	51.51
	0.9	57.97	52.58	51.77	73.05	53.24
	1.0	57.97	52.58	51.77	78.77	54.84
0.99	0.0	104.67	115.10	106.32	104.67	104.67
	0.1	104.67	115.10	106.32	114.91	106.31
	0.2	104.67	115.10	106.32	125.35	108.13
	0.3	104.67	115.10	106.32	135.88	110.08
	0.4	104.67	115.10	106.32	146.37	112.10
	0.5	104.67	115.10	106.32	156.67	114.08
	0.6	104.67	115.10	106.32	166.73	115.97
	0.7	104.67	115.10	106.32	176.56	117.77
	0.8	104.67	115.10	106.32	186.19	119.50
	0.9	104.67	115.10	106.32	195.60	121.16
	1.0	104.67	115.10	106.32	204.74	122.77

Table 5: The estimation error of various estimators in terms of the Matrix 1-Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
0.70	0.0	70.94	36.53	36.82	70.94	70.94
	0.1	70.94	36.53	36.82	55.93	56.28
	0.2	70.94	36.53	36.82	44.71	45.32
	0.3	70.94	36.53	36.82	34.97	35.82

Continued on next page

Table 5: The estimation error of various estimators in terms of the Matrix 1-Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	0.4	70.94	36.53	36.82	26.96	28.05
	0.5	70.94	36.53	36.82	20.38	21.71
	0.6	70.94	36.53	36.82	16.52	17.21
	0.7	70.94	36.53	36.82	14.44	13.75
	0.8	70.94	36.53	36.82	13.20	11.21
	0.9	70.94	36.53	36.82	12.87	10.54
	1.0	70.94	36.53	36.82	12.67	10.18
0.80	0.0	72.33	47.47	46.79	72.33	72.33
	0.1	72.33	47.47	46.79	59.68	60.34
	0.2	72.33	47.47	46.79	48.53	49.80
	0.3	72.33	47.47	46.79	38.78	40.55
	0.4	72.33	47.47	46.79	30.85	32.82
	0.5	72.33	47.47	46.79	25.75	26.66
	0.6	72.33	47.47	46.79	22.88	22.14
	0.7	72.33	47.47	46.79	21.10	18.78
	0.8	72.33	47.47	46.79	19.92	16.11
	0.9	72.33	47.47	46.79	20.04	15.10
	1.0	72.33	47.47	46.79	20.26	14.93
0.90	0.0	81.75	59.18	56.66	81.75	81.75
	0.1	81.75	59.18	56.66	68.10	70.18
	0.2	81.75	59.18	56.66	57.16	61.33
	0.3	81.75	59.18	56.66	48.38	54.47
	0.4	81.75	59.18	56.66	40.41	48.40
	0.5	81.75	59.18	56.66	33.92	43.81
	0.6	81.75	59.18	56.66	33.02	40.06

Continued on next page

Table 5: The estimation error of various estimators in terms of the Matrix 1-Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	0.7	81.75	59.18	56.66	33.81	36.64
	0.8	81.75	59.18	56.66	35.11	33.87
	0.9	81.75	59.18	56.66	36.60	32.95
	1.0	81.75	59.18	56.66	38.18	33.58
0.95	0.0	80.91	79.60	77.41	80.91	80.91
	0.1	80.91	79.60	77.41	75.95	72.89
	0.2	80.91	79.60	77.41	73.70	67.96
	0.3	80.91	79.60	77.41	73.60	64.99
	0.4	80.91	79.60	77.41	73.83	62.35
	0.5	80.91	79.60	77.41	74.22	59.86
	0.6	80.91	79.60	77.41	74.66	57.43
	0.7	80.91	79.60	77.41	75.15	55.09
	0.8	80.91	79.60	77.41	76.05	53.21
	0.9	80.91	79.60	77.41	77.18	52.05
	1.0	80.91	79.60	77.41	78.95	52.05
0.99	0.0	125.39	124.68	138.13	125.39	125.39
	0.1	125.39	124.68	138.13	124.99	122.63
	0.2	125.39	124.68	138.13	124.99	119.93
	0.3	125.39	124.68	138.13	130.98	118.03
	0.4	125.39	124.68	138.13	142.83	116.99
	0.5	125.39	124.68	138.13	154.67	116.69
	0.6	125.39	124.68	138.13	166.26	116.43
	0.7	125.39	124.68	138.13	177.31	117.24
	0.8	125.39	124.68	138.13	188.01	118.44
	0.9	125.39	124.68	138.13	198.11	119.68

Continued on next page

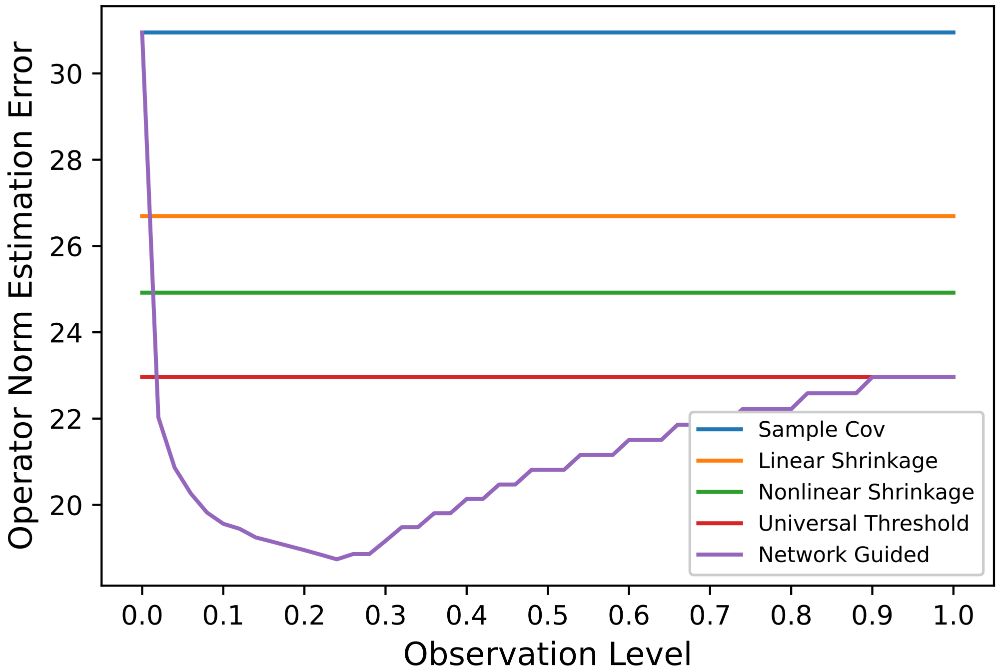

Figure 1: The estimation error against the observation level

Table 5: The estimation error of various estimators in terms of the Matrix 1-Norm

rho	Threshold Level	Sample Cov	Linear Shrinkage	Nonlinear Shrinkage	Universal Threshold	Network Guided
	1.0	125.39	124.68	138.13	207.97	120.93

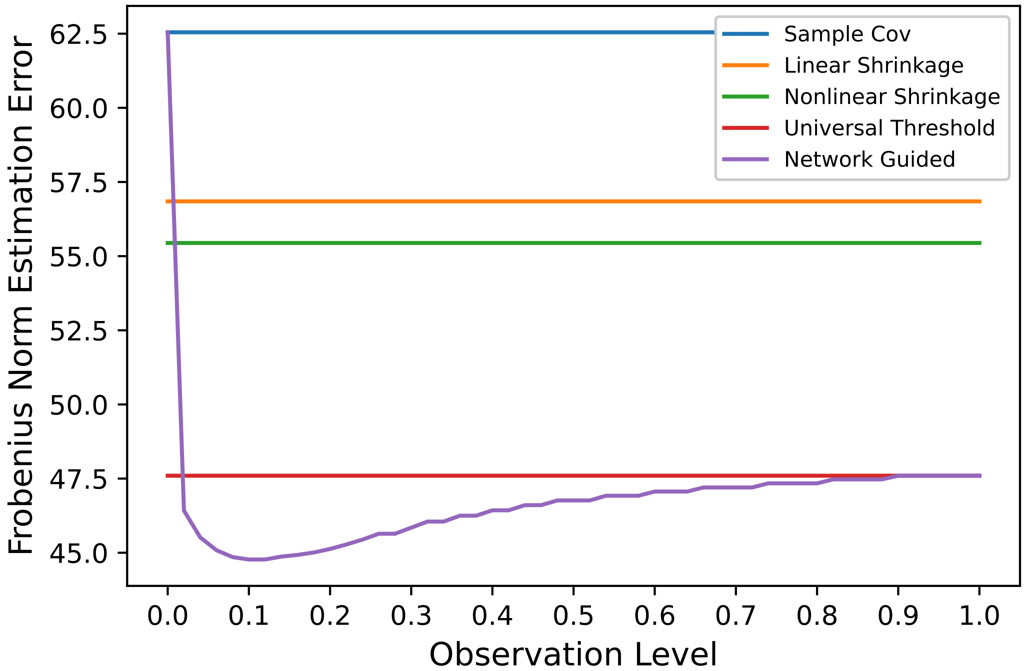

Figure 2: The estimation error against the observation level

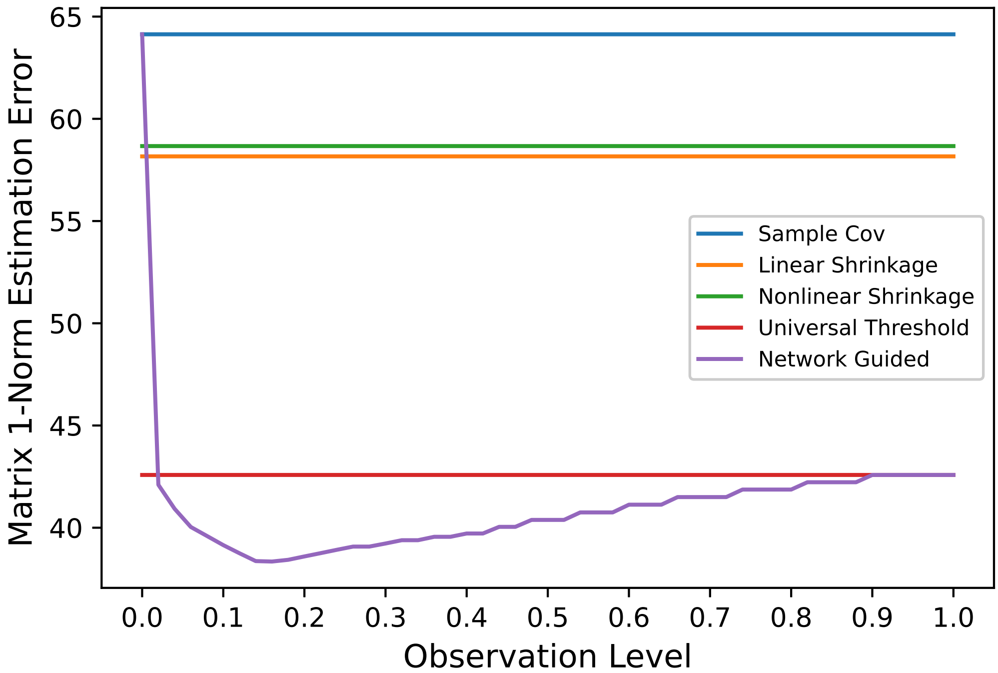

Figure 3: The estimation error against the observation level

Table 6: The estimation error in terms of Operator Norm of the Network Guided Estimator with varying probabilities p, q that determine how G is generated.

q p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	19.74	19.94	20.11	20.49	20.92	21.42	21.81	22.41	22.99	23.62	24.32
0.1	19.35	19.63	19.79	20.17	20.53	21.03	21.41	22.09	22.56	23.30	24.49
0.2	18.98	19.20	19.42	19.78	20.23	20.56	21.11	21.72	22.24	23.46	24.80
0.3	18.68	18.84	19.16	19.49	19.81	20.33	20.70	21.43	22.39	23.80	25.21
0.4	18.18	18.29	18.80	19.03	19.40	19.92	20.30	21.23	22.71	24.17	25.48
0.5	17.87	18.09	18.42	18.73	19.05	19.53	20.38	21.69	23.04	24.57	25.93
0.6	17.48	17.81	18.02	18.45	18.94	19.52	20.63	22.01	23.53	24.91	26.25
0.7	17.15	17.36	17.63	18.09	18.41	19.61	21.00	22.52	23.76	25.19	26.53
0.8	16.80	16.93	17.18	17.64	18.76	20.13	21.36	22.87	24.09	25.54	26.96
0.9	16.42	16.65	16.99	17.84	19.20	20.48	21.84	23.22	24.43	25.92	27.26
1.0	16.07	16.38	16.86	18.22	19.41	20.86	22.11	23.47	24.84	26.28	27.64

Table 7: The estimation error in terms of Frobenius Norm of the Network Guided Estimator with varying probabilities p, q that determine how G is generated.

q p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	47.86	49.31	50.77	52.31	53.56	54.93	56.27	57.55	58.66	59.94	61.17
0.1	47.44	48.86	50.43	51.77	53.18	54.47	55.75	57.02	58.36	59.59	60.82
0.2	47.04	48.54	50.01	51.34	52.68	54.10	55.44	56.80	57.95	59.29	60.48
0.3	46.56	47.93	49.46	51.09	52.33	53.73	55.24	56.36	57.72	58.88	60.12
0.4	46.07	47.50	49.28	50.54	51.96	53.37	54.73	55.90	57.48	58.54	59.82
0.5	45.62	47.13	48.80	50.19	51.59	52.99	54.34	55.71	56.95	58.22	59.36
0.6	45.11	46.75	48.36	49.80	51.13	52.49	54.07	55.22	56.56	57.87	59.10
0.7	44.68	46.24	47.94	49.17	50.66	52.11	53.50	54.98	56.15	57.43	58.72
0.8	44.14	45.85	47.40	48.85	50.29	51.51	53.19	54.46	55.77	57.14	58.33

Continued on next page

Table 7: The estimation error in terms of Frobenius Norm of the Network Guided Estimator with varying probabilities p, q that determine how G is generated.

q p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.9	43.71	45.40	46.94	48.54	49.98	51.35	52.66	54.09	55.52	56.67	57.97
1.0	43.23	44.96	46.41	48.01	49.47	50.98	52.25	53.73	55.04	56.37	57.62

Table 8: The estimation error in terms of Matrix 1-Norm of the Network Guided Estimator with varying probabilities p, q that determine how G is generated.

q p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	46.95	49.00	49.80	52.40	53.62	54.57	57.29	57.80	60.11	63.76	65.74
0.1	46.95	47.53	49.97	51.99	53.87	53.88	55.89	58.53	60.09	62.86	65.74
0.2	46.61	48.19	49.71	49.79	53.39	54.74	56.73	58.43	60.45	62.75	65.16
0.3	45.49	47.53	49.30	50.74	52.88	54.29	56.22	59.85	61.26	61.27	64.46
0.4	45.31	48.02	47.97	49.82	50.84	52.93	56.19	56.59	58.12	61.22	64.03
0.5	45.80	46.56	48.61	49.28	50.60	51.76	54.87	55.77	59.02	61.72	63.86
0.6	44.66	47.15	47.88	49.32	52.90	52.85	54.54	55.43	60.59	61.36	63.56
0.7	44.49	45.75	46.69	49.94	49.30	52.34	54.33	57.37	58.59	60.21	63.63
0.8	44.32	45.04	48.23	48.84	50.29	51.54	54.58	56.66	57.01	60.45	63.01
0.9	43.97	45.09	45.86	48.46	50.47	51.84	54.07	55.14	57.12	59.79	62.65
1.0	43.32	44.17	46.89	47.86	49.40	51.31	53.18	54.54	58.56	59.94	62.07

5.2 Proofs

Proof of Theorem 1. We have the following decomposition:

$$\left\|T_{\hat{L},\lambda}\left(\hat{\Sigma}\right) - \Sigma\right\| \le \left\|T_{\hat{L},\lambda}(\Sigma) - \Sigma\right\| + \left\|T_{\hat{L},\lambda}(\hat{\Sigma}) - T_{\hat{L},\lambda}(\Sigma)\right\| = \mathbf{I} + \mathbf{II}$$

The first term can be bounded by

$$
\begin{aligned}
\mathbf{I} &\leq \max_i \sum_j \left| s_{\hat{L},\lambda}(\sigma_{ij}) - \sigma_{ij} \right| \\
&= \max_i \sum_j \hat{L}^0_{ij} \left| s_\lambda(\sigma_{ij}) - \sigma_{ij} \right| \\
&= \max_i \sum_j \left[L^0_{ij} \left| s_\lambda(\sigma_{ij}) - \sigma_{ij} \right| + (\hat{L}^0_{ij} - L^0_{ij}) \left| s_\lambda(\sigma_{ij}) - \sigma_{ij} \right| \right] \\
&\leq (1 + o_p(1)) \max_i \sum_j \left[L^0_{ij} \left| s_\lambda(\sigma_{ij}) - \sigma_{ij} \right| \right] \\
&\leq (1 + o_p(1)) \max_i \sum_j \left[L^0_{ij} \left| \sigma_{ij} \mathbf{1}\{\sigma_{ij} \leq \lambda\} + (s_\lambda(\sigma_{ij}) - \sigma_{ij}) \mathbf{1}\{\sigma_{ij} > \lambda\} \right| \right] \\
&\leq (1 + o_p(1)) \max_i \sum_j \left[L^0_{ij} \left| \sigma_{ij} \right|^q \lambda^{1-q} \right] \\
&\leq (1 + o_p(1)) c_0(p) \lambda^{1-q}
\end{aligned}
$$

And the second term can be bounded similar to Rothman, Levina, and Zhu, 2009,

$$
\begin{aligned}
\mathbf{II} &\leq \max_i \sum_j \left[\hat{L}^1_{ij} \left| \hat{\sigma}_{ij} - \sigma_{ij} \right| + \hat{L}^0_{ij} \left| s_\lambda(\hat{\sigma}_{ij}) - s_\lambda(\sigma)_{ij} \right| \right] \\
&\leq (1 + o_p(1)) c_1(p) \max_{ij} \left| \hat{\sigma}_{ij} - \sigma_{ij} \right| + (1 + o_p(1)) \max_i \sum_j L^0_{ij} \left| s_\lambda(\hat{\sigma}_{ij}) - s_\lambda(\sigma)_{ij} \right| \\
&= O_p\left(c_1(p) \sqrt{\frac{\log N}{T}} + c_0(p) \left(\lambda^{1-q} + \lambda^{-q} \sqrt{\frac{\log N}{T}} \right) \right)
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\left\| T_{\hat{L},\lambda}\left(\hat{\Sigma}\right) - \Sigma \right\| &= O_p\left(c_1(p) \sqrt{\frac{\log N}{T}} + c_0(p) \left(\lambda^{1-q} + \lambda^{-q} \sqrt{\frac{\log N}{T}} \right) \right) \\
&= O_p\left(c_1(p) \sqrt{\frac{\log N}{T}} + c_0(p) \left(\frac{\log N}{T} \right)^{\frac{1-q}{2}} \right)
\end{aligned}
$$

□

5.3 Data Description: News Implied Network

The news data are obtained from RavenPack Equity files Dow Jones Edition for the period January 2004 to December 2015. This comprehensive news dataset combines relevant content from multiple sources, including Dow Jones Newswires, Wall Street Journal, and Barron's MarketWatch, which produce the most actively monitored streams of news articles in the financial system. Each unique news story (identified by a unique story ID) tags the companies mentioned in the news by their unique and permanent entity identifier codes (RP_ENTITX_ID), by which we link to stock identifier TICKER and PERMNO.

As as Ge and O. B. Linton, 2021, we identify links by news co-mentioning. That is, if a piece of business news reports two companies together, they share a link. We do not consider news that co-mention more than two companies since although news they may carry potential information about links, they provide noisier information. We also remove news with topics including analyst recommendations, rating changes, and index movements as these types of news might stack multiple companies together when they actually do not have real links. Table 9 provides descriptive statistics for RavenPack Equity files Dow Jones Edition dataset during the sample period. Since our comprehensive news dataset combines several sources, given a similar length of sample period, the number of unique news stories is more than ten times larger than that from Scherbina and Schlusche, 2015 and more than eight hundred times than that from Schwenkler and Zheng, 2019. For link identification purposes, we only use sample news (1) are not about topics mentioned above (2) tag *S&P* 500 companies and (3) mention exactly two companies, which is a subsample of $1,637,256$ unique news stories.

Number of unique news stories	$88,316,898$
Number of stories remaining after removing topics including analyst recommendations, ratings changes, and index movements	$87,841,641$
Of these:	
Number of stories tag sample companies	$8,341,848$
Of these:	
Number of stories that mention only one company	$5,507,772$ (66.03%)
Number of stories that mention exactly two companies	$1,637,256$ (19.63%)
Number of stories that mention more than two companies	$1,196,820$ (14.34%)

Table 9: Descriptive statistics for RavenPack Equity files Dow Jones Edition for the period January 2004 to December 2015.

References

Bickel, Peter J. and Elizaveta Levina (2008a). "Covariance Regularization by Thresholding". In: *The Annals of Statistics* 6, pp. 2577–2604 (cit. on pp. 3, 5).

— (2008b). "Regularized Estimation of Large Covariance Matrices". In: *The Annals of Statistics* 1, pp. 199–227 (cit. on p. 2).

Brownlees, Christian, Gumundur Stefán Gumundsson, and Gábor Lugosi (2020). "Community detection in partial correlation network models". In: *Journal of Business & Economic Statistics*, pp. 1–11 (cit. on pp. 3, 4).

Cai, Tony and Weidong Liu (2011). "Adaptive Thresholding for Sparse Covariance Matrix Estimation". In: *Journal of the American Statistical Association* 494, pp. 672–684 (cit. on pp. 3, 6).

Chen, Jia, Degui Li, and Oliver Linton (2019). "A new semiparametric estimation approach for large dynamic covariance matrices with multiple conditioning variables". In: *Journal of Econometrics* 212.1, pp. 155–176 (cit. on p. 11).

Fan, Jianqing, Alex Furger, and Dacheng Xiu (2016). "Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data". In: *Journal of Business & Economic Statistics* 34.4, pp. 489–503 (cit. on p. 3).

Fan, Jianqing, Yuan Liao, and Han Liu (Apr. 12, 2015). *An Overview on the Estimation of Large Covariance and Precision Matrices*. arXiv: 1504.02995 [stat]. URL: http://arxiv.org/abs/1504.02995 (cit. on pp. 3, 10).

Fan, Jianqing, Yuan Liao, and Martina Mincheva (2013). "Large covariance estimation by thresholding principal orthogonal complements". In: *Journal of the Royal Statistical Society. Series B, Statistical methodology* 75.4 (cit. on p. 3).

Ge, Shuyi and Oliver B Linton (2021). "News-Implied Linkages and Local Dependency in the Equity Market". In: *Available at SSRN 3827902* (cit. on pp. 3, 4, 25).

Hoberg, Gerard and Gordon Phillips (2016). "Text-based network industries and endogenous product differentiation". In: *Journal of Political Economy* 124.5, pp. 1423–1465 (cit. on p. 3).

Kaustia, Markku and Ville Rantala (2013). "Common analyst-based method for defining peer firms". In: *Available at SSRN* (cit. on p. 3).

Ledoit, Olivier and Michael Wolf (Feb. 1, 2004a). "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices". In: *Journal of Multivariate Analysis* 2, pp. 365–411. ISSN: 0047-259X. DOI: 10.1016/S0047-259X(03)00096-4 (cit. on p. 3).

— (July 31, 2004b). "Honey, I Shrunk the Sample Covariance Matrix". In: *The Journal of Portfolio Management* 4, pp. 110–119. ISSN: 0095-4918, 2168-8656. DOI: 10.3905/jpm.2004.110 (cit. on pp. 3, 9).

— (Apr. 2012). "Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices". In: *Annals of Statistics* 2, pp. 1024–1060. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/12-AOS989 (cit. on p. 3).

— (Dec. 1, 2017). "Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks". In: *The Review of Financial Studies* 12, pp. 4349–4388. ISSN: 0893-9454, 1465-7368. DOI: 10.1093/rfs/hhx052 (cit. on p. 3).

Rothman, Adam J., Elizaveta Levina, and Ji Zhu (2009). "Generalized Thresholding of Large Covariance Matrices". In: *Journal of the American Statistical Association* 485, pp. 177–186 (cit. on p. 24).

Scherbina, Anna and Bernd Schlusche (2015). "Economic linkages inferred from news stories and the predictability of stock returns". In: *Available at SSRN 2363436* (cit. on pp. 4, 25).

Schwenkler, Gustavo and Hannan Zheng (2019). "The network of firms implied by the news". In: *Available at SSRN 3320859* (cit. on pp. 4, 25).

Shu, Hai and Bin Nan (June 1, 2019). "Estimation of Large Covariance and Precision Matrices from Temporally Dependent Observations". In: *The Annals of Statistics* 3. ISSN: 0090-5364. DOI: 10.1214/18-AOS1716. arXiv: 1412.5059 (cit. on pp. 3, 6).