Financial factors selection with knockoffs: fund replication, explanatory and prediction networks

Damien Challet, Christian Bongiorno, Guillaume Pelletier

Version of Record: https://www.sciencedirect.com/science/article/pii/S0378437121003782
Manuscript_a105a82a4182096b0c94dd51ad27ef69

Financial factors selection with knockoffs: fund replication, explanatory and prediction networks

Damien Challet[1], Christian Bongiorno[1], and Guillaume Pelletier[2]

[1] *Université Paris-Saclay, CentraleSupélec, Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, 91192 Gif-sur-Yvette, France*
[2] *Automated Market Making, BNP Paribas, 20 boulevard des Italiens, 75009 Paris, France*

Abstract

We apply the knockoff procedure to factor selection in finance. By building fake but realistic factors, this procedure makes it possible to control the fraction of false discovery in a given set of factors without resorting to p-values. To show its versatility, we apply it to fund replication and to the inference of explanatory and prediction networks.

Keywords: factors, factor selection, clustering, lead-lag, prediction, backtest

1. Dedication

At a time when multi-processor desktop PCs were a rarity, there was a special breed of physicists that had mastered massively parallel computing techniques. Dietrich Stauffer was one of them. He also was a parallel researcher, working simultaneously on many topics and papers. This made a durable impression on the young researcher that I (DC) was then.

This submission required massively parallel computations (600 CPU cores for a few days). Dietrich would not have been impressed.

2. Introduction

Factor hunting is an age-old task in Finance, either in an explanatory pursuit, or in a prediction setting. Factor may explain risk, performance and diversification. This endeavour can be divided into two parts: finding candidate factors and selecting them.

The celebrated Fama-French factors [10, 11] have two precious qualities: they are exactly a handful nowadays (5, up from 3 [11]) and make sense from a financial point of view. Their small number is a sure way to avoid data overfitting with too many factors, and their financial interpretation is straightforward in layman's terms. On the other hand, it is beyond doubt that they cannot possibly capture all the subtleties of price return dynamics and that additional factors, possibly ephemeral ones, are needed.

Dropping the interpretability requirement, it is nowadays very easy to generate a large quantity of candidate factors from fundamental or alternative data. Selecting them becomes a difficult task. Statistics suggests to regard factor selection among a given set of factors as multiple hypotheses testing: to each factor corresponds a null hypothesis that it is irrelevant. This opens the way to methods that are able to control the error rate when one selects factors [4]. Trying to keep only relevant factors being unrealistic, one settles for a less ambitious aim. In finance in particular, one can tolerate a controlled fraction of wrong choices given the amount of noise contained in the data.

Whereas the usual methods of controlling the error rate are based on p-values, Ref. [1] introduced the so-called knockoff procedure which dispenses with p-values altogether. It consists in creating a fake but look-alike factor for each candidate factor, i.e., a knockoff (a copy of low quality) and to add all knockoffs to set of candidate factors. Because one can labels at least some factors as irrelevant, one can control the error rate in any selection. Practically, one ranks all these factors according to some relevance metrics and the knockoff procedure finds the relevance threshold that controls the fraction of false factors that have a relevance larger than this threshold. Since its introduction, this method has spanned a flurry of new methods and has been steadily improved (see Refs. [7, 16, 12, 25] for example). Here we apply them to raw financial asset price returns in three situations: fund replication, explanatory and prediction networks.

3. The selection problem

Let us introduce some useful notations. Here, we are mostly interested in explaining the price returns of a given asset, denoted by r from a matrix of N candidate factors, denoted by R (which may include a shifted r in a predictive context). In a regression setting, one writes

$$r = R\beta + z \quad \text{with} \quad \begin{cases} r \in \mathbb{R}^T \\ R \in \mathbb{R}^{N\times T} \end{cases}, \tag{1}$$

where $\beta \in \mathbb{R}^N$ are the factor loadings and $z \in \mathbb{R}^T$ the residuals. In this paper, the candidate factors only consist of price returns of a collection of assets. More generally, one can write $r = F(R) + z$ where F is a non-linear function, for example a Random Forest [6].

A given factor i is selected whenever $\beta_i \neq 0$. It is highly likely that a least-squares optimization yields $\beta_i \neq 0$ for all values of i even if $N \gg T$, which results in overfitting. A well-known way to restrict the number of selected factors is to add an $\mathbb{L}_1$ penalization to the least-square problem. The LASSO [27] in particular can be written as

$$\hat{\beta}(\lambda) = \underset{\beta\in\mathbb{R}^N}{\operatorname{argmin}} \left\{ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}, \tag{2}$$

where the penalization constant λ controls the number of selected components, i.e., non-null elements of $\hat{\beta}$. More precisely, each factor i is selected when $\lambda < \lambda_i$; intuitively, relevant factors should be selected for larger values of λ than irrelevant ones. Thus choosing a value of λ yields a selection of factors $S(\lambda) = \{i \text{ s. th. } \lambda_i < \lambda\}$. The whole question is how to choose λ and according to which criterion. Fixing the number of selected factors does not offer any statistical guarantee.

Let us denote the set of index of the relevant factors, or equivalently, the selection set of the true factors by $S \subset \{1, \cdots, N\}$ and a given factor selection (obtained by whatever method) by $\hat{S}$ such that such $i \in \hat{S} \iff \hat{\beta}_i \neq 0$ and inversely. The False Discovery Proportion (FDP henceforth) of a given selection $\hat{\beta}$ is defined by:

$$\text{FDP} = \frac{\#\{j : j \in \hat{S} \setminus S\}}{\#\{j : j \in \hat{S}\}}, \tag{3}$$

with the convention that FDP = 0 when no factor is selected. By definition, the False Discovery Rate is FDR = E(FDP). Then, a selection rule controls the FDR at level q if FDR $\leq q$. Note that this does not control the false rejection rate, i.e., the falsely rejected factors. The canonical way to control the FDR rests on correcting the threshold of p-values (see [3] for a review).

3.1. Knockoffs

Ref. [1] proposes a completely different approach to solve the problem of controlling the FDR of selected factors by adding N irrelevant but specially crafted factors to R. This means that any factor ranking method yields a mixture of some factors known to be irrelevant and some possibly relevant. This, in essence, makes it possible to control the FDR without resorting to p-values. The power of knockoffs is similar or better to the canonical FDR control way according to [1].

More specifically, the quality of each factor is assessed with some method, which essentially yields a statistics (a real number) denoted by Z; we assume that a larger Z is more likely to be associated with a relevant factor. Let us denote by $\tilde{X}$ the matrix of knockoff factors, and Z_i and $\tilde{Z}_i$ the quality statistics of factor i and of its knockoff. Clearly, both distributions of Z_i and $\tilde{Z}_i$ are the same if i is irrelevant, while one expects that $E(Z_i) > E(\tilde{Z}_i)$ otherwise. Then the FDR is controlled according to the fraction of candidate factors such that $Z_i < \tilde{Z}_i$ in a given selection set $\hat{S}$. For more details, the reader is referred to [1, 7].

For example, in the context of the LASSO problem, on average, relevant factors will be selected for larger values of λ than their knockoffs, while there is no such ordering for irrelevant factors. Thus Z_i can be the maximal value of λ such that factor i has $\beta_i \neq 0$. Other possibilities include the feature importance of factor i in tree-based machine learning methods, which will be used later in this paper.

There are several ways to build knockoffs. Here, we use the so-called model-X method [7], which can be used in the high dimensional case $N > T$. It rests on two assumptions: r is independent from the irrelevant features and the distribution of $[R, \tilde{R}]$ is invariant under the exchange of a factor and its knockoff.

Constructing knockoffs that respect these two conditions is generally hard. For the sake of computational speed, we use an approximate second order moment matching method from the `knockoff R` package [24]. More powerful methods, based on deep learning [19, 25, 26] and reversible Markovian chains [2], deal better with heavy tailed data and more complex dependencies. Given its additional computational burden and the already considerable computation power needed to produce our results, we leave the inclusion of these methods to a future study.

4. Results

4.1. Fund replication

Among some notable applications of this method to finance, fund replication is obtained by regressing the performance of the latter to the returns of strategies that it may potentially use, given the constraints of its prospectus (see e.g. [30]).

A straightforward application of knockoffs is to find out what sparse combination of assets can explain (and reproduce) the price returns of a given fund, r_t. We consider returns between adjusted daily close prices of 4400 US equities in the 2005-2016 period, with a calibration window of 252 days (about a year). In each calibration window, we remove assets which have any missing value. Even if the knockoff model-X method allows for having more assets than timesteps, $T = 252$ and $N = 4400$ would be quite ambitious. Thus, we perform 200 sampling without replacement of 500 assets each, which is still in the high dimensional regime.

As a simple illustration, we select an ETF which attempts to replicate the performance of the energy sector, XLE. For each calibration window, we compute the ratio between the number of selected assets that belong to the energy sector according to the GICS classification [20] and the total number selected assets. In principle, if there are no correlations between the energy sector ETF and assets not classified as in the energy sector, this ratio should equal $1 - \mathrm{q}$, where q is the chosen FDR level. We ran computations for various values of q to check that knockoffs are consistent. Figure 1 reports that the realized FDR is close to q for $q \in [0.12, 0.35]$. We note that in principle, the knockoffs guarantee that the realized FDR is smaller than q. The problem here is that financial assets are not independent from each other, thus that the returns of XLE may be explained by many other tickers that do not belong in the energy sector. For example, the average correlation between XLE and Ford (F) in the $2007-2021$ period is about 0.5. As a consequence, a false positive (according to the industrial sector) does not qualify as a false positive from a statistical point of view. This explains the increase of realized FDR for $q \geq 0.4$, as assets with decreasing correlation with XLE, thus more likely not to belong in the

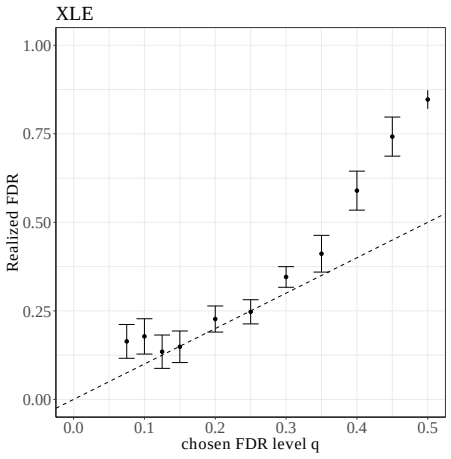

Figure 1: Realized FDR versus chosen FDR when replicating the performance of XLE, an energy sector ETF. 200 samples of 500 assets, 252 days of calibration length, average over 12 years.

energy sector, are included in the selection made by the knockoff procedure. The discrepancy for small FDR comes from the fact that the typical number of assets belonging to the energy sector selected by the knockoff procedure reaches 1 and is therefore a finite investment universe effect. In short, this simple test case confirms that knockoffs give meaningful results in a context in which the success rate is roughly testable.

4.2. Explanatory networks

Knockoffs can be used to build statistically validated explanatory networks of price returns, where there is a *directed* link from asset j to asset i ($i \neq j$) if $r_{j,t}$ contributes to explain $r_{i,t}$, Let us denote by $N_i^{(t_0,t_1)}$ the set of assets that explain the price returns of asset i in a given calibration window $[t_0, t_1]$ at a given FDR level. In contrast to many other clustering methods based on symmetric measures (e.g. correlations [18, 15] or anomalous synchronization as in Statistically Validated Networks (SVNs) [29]), knockoff explanatory networks are directed: if asset 2 and 3 explain asset 1, it may happen that asset 1 does not belong to $N_2^{(t_0,t_1)}$. This implies that the knockoffs explanatory networks are more flexible than correlation-based methods and may contain more information. The downside is that more work is needed to extract clusters from these results, for example by performing network clustering, known as community detection in this stream of literature (see [13] for a review).

Here we focus on the time evolution of network properties especially between industries as defined by the GICS classification. Because computing time scales as N^2 where N is the number of assets to test, we focus on 200 US equities from 2000 to the end of 2017, which leads to reasonable computing time on several

hundred CPU cores. We repeat the knockoff generation and selection process a given number of times (100 here) for each asset and for each calibration window and consider the union of all the selected factors. This is because the selection is empty for a sizable fraction of times, a known instability of knockoff generation (see e.g. [16]). We found that the knockoff selection using feature importance of candidate factors and knockoffs determined by random forests produced the most factors at a given level of FDR.

Figure 2 reports the time evolution of four network synthetic measures. The link density, defined as the number of observed links divided by the all possible directed links ($N(N-1)$), is relatively high on average and has clear dips during the 2008-2009 and subsequent crises. To understand the origin of this phenomenon, we plot the reciprocity, i.e., the fraction of links that are bidirectional, which shows a similar behavior. The assortativity of links between sectors, which measures the propensity to establish links between nodes of the same sector (see [21]) is clearly significant, which is not surprising. It is noteworthy that it behaves in an opposite way from reciprocity [22]. During times of crisis, links from the same industrial sectors are more likely to remain relevant.

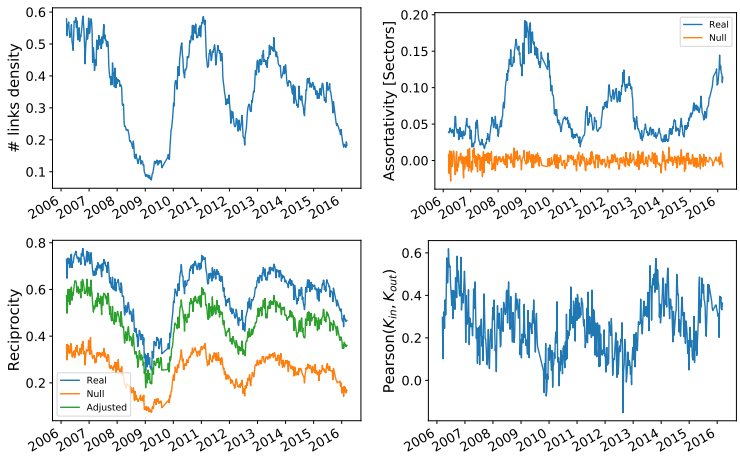

Figure 2: Upper left panel shows the link density; upper right panel shows the assortativity of links between sectors and assortativity of a configuration model that preserves the in- and out-degree of the nodes; lower left panel shows the link reciprocity, the figure also reports the reciprocity of a configuration model that preserves in and out degree of the node and the adjusted estimator [14]; lower right panel shows the Pearson correlation between the in-degree and out-degree of the nodes. FDR= 0.1, calibration windows of $T_{in} = 300$ timesteps.

4.3. Prediction networks

Using knockoffs to explain current returns by lagged ones yields statistically validated prediction networks, as shown in Fig 3. Such networks are directed: for each asset i and time t, a collection of predictive assets $P_{i,t}$ links to i. We note that the in- out-links degrees is asymmetric: the number of out-going links is

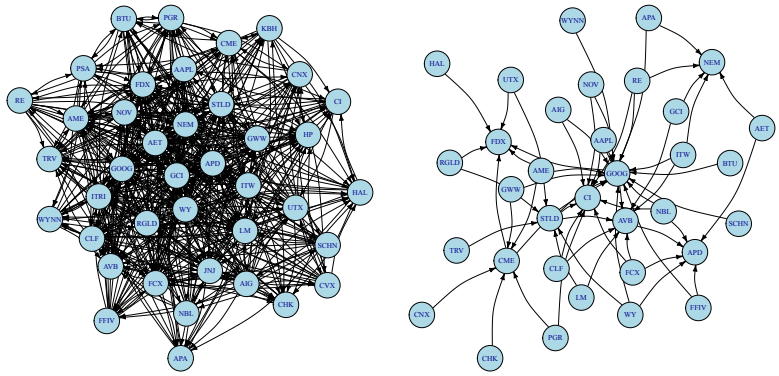

Figure 3: Examples of prediction network inferred with knockoffs. $T_{in} = 300$ trading days, period ending on 2006-03-13. Left plot: FDR= 0.3 and 100 runs per asset; right plot: FDR= 0.2 and 10 runs per asset. Selection via random forest variable importance.

usually much smaller than the incoming links. What we call prediction networks here are also called lead-lag networks, obtained for example by Granger causality with multiple hypothesis correction (see e.g. [23] for a recent example among many), by the asymmetry of lagged correlations [28, 17, 9], or by extending SVNs to lagged time series [8]. The difference here is that we do not need to make any assumption about the Gaussianity of price returns as in Granger causality, and that we do not use a correction on p-values: the ranking of factors is obtained by Random Forests, an inherently non-linear and distribution agnostic method. Note also that the method here also works in the high-dimensional regime (as do SVNs).

We use exactly the same numerical setup as in the explanatory networks part except for the lag of the predictors. Figure 3 shows two examples of prediction networks obtained in the same calibration window, but with a different level FDR. The influence of that choice on the link density is very large: it is much more difficult to control an FDR set to 0.2 than 0.3 and hence much fewer links are found; almost no links are found for FDR= 0.1, which means that it is almost impossible to guarantee a small FDR in a price return prediction context. However, it does not matter too much: as long as method is able to extract some information, it will still yield a predictive power. However, the larger the FDR, the more the predictions will be potentially affected by noise.

At time t, for each asset i such that $P_{i,t} \neq \emptyset$, we apply a robust linear fit of $r_{i,t+1}$ with real returns of $P_{i,t}$ over the last T_{in} days in-sample, and then predict the out-of-sample return for the next $\delta T = 5$ days from the previous daily returns of each predictive assets. We first check the hit ratio of the sign of the prediction as a function of the number of elements in $P_{i,t}$, denoted by k_{in} (Fig. 4): quite remarkably, it decreases as a function of k_{in} in a similar way for FDR= 0.2 and 0.3.

While prediction networks can be used to build a trading strategy, the performance we report in the following cannot be considered as a proper back-test,

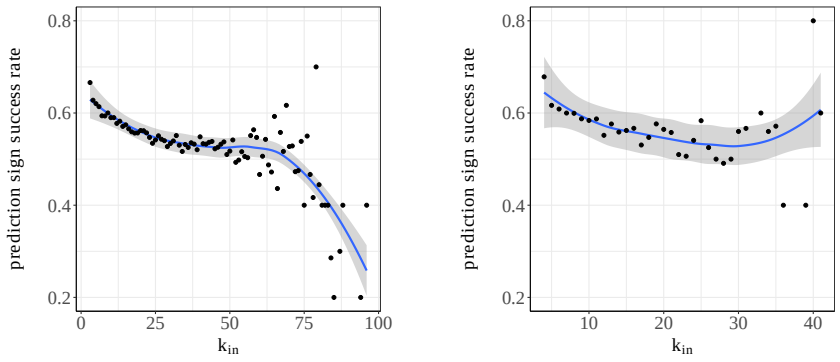

Figure 4: Prediction hit ratio (out-of-sample) as a function of the number of assets k_{in} that belong in predictors' set. Left plot: FDR=0.3; right plot: FDR= 0.2. $T_{in} = 300$; 200 assets. Blue curve: local average; gray area: standard error on the local average.

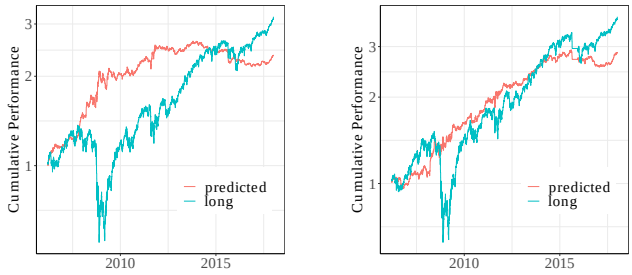

Figure 5: Cumulative performance from prediction networks computed for predicted assets with equally weighted portfolios either long-short (predicted), or long-only (long). $T_{in} = 300$ Left plot: FDR= 0.3; right plot FDR= 0.2.

as we use closing price both for computing returns and to open virtual positions, and because we do not include transaction costs. The point of this section is to show the information gain provided by filtering factors with knockoffs, not to suggest a trading strategy. We thus ran two experiments to assess the out-of-sample performance of the predicted assets: first we take the predicted assets, take a position according to the signs of their predicted returns, and compute the performance of equally-weighted portfolios, rebalanced at every timestep; in order to compare the information contained in the sign of the returns, we also add the performance of equally-weighed portfolios of long positions on predicted assets, which gives a performance similar to that of the market. Both levels of FDR have acceptable performance before a saturation later on (2012-2017); this is in part because they provide long-short predictions, which worked better before 2015. We note however that FDR= 0.2 leads to a better performance before 2015, as the signal provided by more trustworthy (and fewer) factors is better.

We then use the knockoffs selection and predictors in a more subtle way:

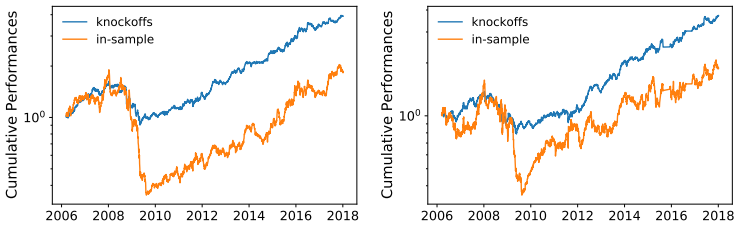

Figure 6: Cumulative performance from prediction networks with mean-variance portfolios with target daily return equal to 0.005, blue curve uses the knockoff prediction for the returns, the orange curve uses the average in-sample returns over the past 300 days. Left plot: FDR= 0.3; right plot FDR= 0.2.

for each day, we first compute the covariance matrix of all predicted assets over the last T_{in}, filtered with the BAHC method [5]. We then compute optimal long-short mean-variance portfolios at each time step with a fixed net leverage of 1. An interesting variation consists in using the predicted returns as the expected returns instead of using the historical averaged returns. Figure 6 shows the information gain provided by predicted returns. The difference of the gain profile with respect to equally weighted portfolios of Fig. 5 comes from the fact that fixing a net leverage to 1 induces a bias towards long positions, which is detrimental in bear markets.

5. Conclusion

Factor selection with knockoffs holds many promises in finance. This contribution only skims the surface by using price returns as factors. The originality of the knockoff method is that they define directed networks where the presence of links is statistically controlled. Further work includes a better way to generate knockoffs, knockoffs with proper covariance matrix cleaning and applying knockoffs to other kinds of factors.

6. Acknowledgements

This publication stems from a partnership between CentraleSupélec and BNP Paribas.

This work was performed using HPC resources from the “Mésocentre” computing center of CentraleSupélec and École Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France.

References

[1] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[2] S. Bates, E. Candès, L. Janson, and W. Wang. Metropolized knockoff sampling. *Journal of the American Statistical Association*, pages 1–15, 2020.

[3] Y. Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society: series B (Methodological)*, 72(4):405–416, 2010.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300, 1995.

[5] C. Bongiorno and D. Challet. Covariance matrix filtering with bootstrapped hierarchies. *PLOS ONE*, 16(1):e0245092, 2021.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: series B (Methodology)*, 80(3):551–577, 2018.

[8] D. Challet, R. Chicheportiche, M. Lallouache, and S. Kassibrakis. Statistically validated lead-lag networks and inventory prediction in the foreign exchange market. *Advances in Complex Systems*, 21(08):1850019, 2018.

[9] C. Curme, M. Tumminello, R. N. Mantegna, H. E. Stanley, and D. Y. Kenett. Emergence of statistically validated financial intraday lead-lag relationships. *Quantitative Finance*, 15(8):1375–1386, 2015.

[10] E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[11] E. F. Fama and K. R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

[12] Y. Fan, J. Lv, M. Sharifvaghefi, and Y. Uematsu. Ipad: stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association*, 115(532):1822–1834, 2020.

[13] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5): 75–174, 2010.

[14] D. Garlaschelli and M. I. Loffredo. Fitness-dependent topological properties of the world trade web. *Physical Review Letters*, 93(18):188701, 2004.

[15] L. Giada and M. Marsili. Data clustering and noise undressing of correlation matrices. *Physical Review E*, 63(6):061101, 2001.

[16] J. R. Gimenez and J. Zou. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2184–2192. PMLR, 2019.

[17] N. Huth and F. Abergel. High frequency lead/lag relationships—empirical facts. *Journal of Empirical Finance*, 26:41–58, 2014.

[18] L. Kullmann, J. Kertesz, and R. Mantegna. Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions. *Physica A: Statistical Mechanics and its Applications*, 287(3-4):412–419, 2000.

[19] Y. Liu and C. Zheng. Auto-encoding knockoff generator for FDR controlled variable selection. *arXiv preprint arXiv:1809.10765*, 2018.

[20] MCSI. The Global Industry Classification Standard. `https://www.msci.com/gics`. Last accessed: 2021-02-15.

[21] M. Newman. *Networks: an introduction*, page 225. Oxford University Press, 2010.

[22] M. E. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

[23] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks. Financial networks based on Granger causality: A case study. *Physica A: Statistical Mechanics and its Applications*, 482:65–73, 2017.

[24] E. Patterson and M. Sesia. *knockoff: The Knockoff Filter for Controlled Variable Selection*, 2020. URL `https://CRAN.R-project.org/package=knockoff`. R package version 0.3.3.

[25] Y. Romano, M. Sesia, and E. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

[26] M. Sudarshan, W. Tansey, and R. Ranganath. Deep direct likelihood knockoffs. *arXiv preprint arXiv:2007.15835*, 2020.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: series B (Methodological)*, 58(1):267–288, 1996.

[28] B. Tóth and J. Kertész. Increasing market efficiency: Evolution of cross-correlations of stock returns. *Physica A: Statistical Mechanics and its Applications*, 360(2):505–515, 2006.

[29] M. Tumminello, S. Micciche, F. Lillo, J. Piilo, and R. N. Mantegna. Statistically validated networks in bipartite complex systems. *PLOS ONE*, 6 (3):e17994, 2011.

[30] V. Weber and F. Peres. Hedge fund replication: Putting the pieces together. *Available at SSRN 2202270*, 2013.