Stock Network Investment
An Application to the Brazilian Stock Market

David Cecchini[a], 黄欣欣[b]

[a]TBSI - Environmental Sciences and New Energy Technology - 2019380040
[b]TBSI - 精准医学与公共健康- 2019214661

June, 2020

Contents

Chapter 1

Background

1.1 Brazil Stock Exchange and Over-the-Counter Market

Brasil, Bolsa, Balcão - B3 (Brazil, Exchange, Counter)[1] is the biggest Brazilian exchange among the top exchanges by market cap in the world, ranking number 18[2], with BRL 4 billion in capitalization (approximately USD 660 billion, value that changes considerably due to fluctuations of dollar to Brazilian real conversion rates) and 330 listed companies.

B3 is a fusion of traditional exchanges in Brazil (Sao Paulo Stock Exchange, Rio Stock Exchange, Brazilian Mercantile and Futures Exchange - BM&F) and *CETIP* (Central of Custody and Financial Settlement of Securities) to form the unified Brazilian exchange.

1.2 Behavioral Finance Hypothesis

Behavioral economics theory studies the limit from rational and irrational decisions made by economic agents. It is known that due to the psychological differences of the agents, they may behave in an irrational way, overreacting or underreacting to market changes.

Based on the assumption that the traders' irrational decisions can cause mispricing to financial assets, investors can design trading strategies that take advantage of the mispriced assets and invest according to the real (fair) price to guarantee a positive return on investments. This kind of investment strategy is called financial behavioral investment.

Portfolio selection is an important part of the investor's decisions since there are many different assets to invest in, each one with different expected returns and different risks. Modern Portfolio Theory (MPT)[3] uses a mathematical approach to select stocks based on the duality *Risk-Return*. The MPT was introduced by the Nobel prize Harry Markowitz[8], where he introduces the concept of diversification that allows a portfolio to obtain similar or higher returns with less risk by adding assets to it.

In this project, we implement a portfolio selection strategy based on a stock network[12]. The strategy is similar to MPT in the sense that it assesses the portfolio risk-return and selects the stocks that minimize the risk, and by doing so we would obtain better returns.

[1] http://www.b3.com.br
[2] https://en.wikipedia.org/wiki/List_of_stock_exchanges
[3] https://en.wikipedia.org/wiki/Modern_portfolio_theory

1.3 Trading premises

In Brazil, securities are processed *B3* and regulated by the Securities Commission of Brazil (CVM) that is independent but directly linked with the Brazilian Ministry of Finance. It regulates markets such as the stock exchange, financial intermediaries, and public companies.

Since March 30, 2017, the Brazilian stock market is unified at *B3*. Before that, most of the companies were listed at the Sao Paulo Stock Exchange (*Bovespa*). The transaction premise has not changed during the period of this strategy, and it can be expected that the transaction premise of Brazil's financial market will not change in the short term.

Chapter 2

The strategy

2.1 Description

Stock prices have a complex dynamics and their movements depend on many factors from economic, financial, and behavioral aspects of the market and it's agents. Our strategy aims to identify correlations between stocks and select the best for a long term investment. In short, we will use a special kind of correlation metric that is suitable for time series and build a network by connecting the stocks based on their "loss-spreading factor" (how the stocks are correlated, and to how many other stocks they are correlated).

Financial time series forecasting is a very complex problem, and Pearson's correlation may not be appropriate to measure the dependencies between the stocks because it detects only linear relations, not to mention that it can have a value equal to zero when the series are dependent[11]. We will use the Distance Correlation metric[11, 10], which is a correlation measure that captures both linear and non-linear relations in the data. From its definition, it also allows time series with different dimmension to be compared but we will study only stocks in the same period in this study. According to [11], the distance covariance is defined as: Let X, Y be two real-valued random variables (vectors) and $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a n-size sample. Then, we first compute the pairwise distances for all $j, k \in \{1, \ldots, n\}$:

$$a_{j,k} = \|X_j - X_k\| \tag{2.1}$$
$$b_{j,k} = \|Y_j - Y_k\| \tag{2.2}$$

Then, define the matrices:

$$A_{j,k} = a_{j,k} - \bar{a}_{j,\cdot} - \bar{a}_{\cdot,k} + \bar{a}_{\cdot,\cdot} \tag{2.3}$$
$$B_{j,k} = b_{j,k} - \bar{b}_{j,\cdot} - \bar{b}_{\cdot,k} + \bar{b}_{\cdot,\cdot} \tag{2.4}$$

where $\bar{a}_{\cdot,k}$ is the k-th columns mean, $\bar{a}_{j,\cdot}$ is the j-th row mean and $\bar{a}_{\cdot,\cdot}$ is the total mean. Same notation for matrix B. Then, the distance covariance is defined by:

$$dCov_n^2(X, Y) := \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} A_{j,k} B_{j,k} \tag{2.5}$$

And the correlation, as usual, is defined by:

$$dCor(X,Y) := \frac{dCov(X,Y)}{\sqrt{dCov(X,X)dCovY,Y}} \tag{2.6}$$

With the distance correlation between all assets computed, we can define our weighted stock network using the winner-take-all method [12]. This method defines the edges weight matrix using a threshold value ρ_c. We want this hyperparameter to be "big enough" (we want to limit the number of connections between low correlated stocks) but at the same time not too big so that the graph remains irreducible (fully connected). The final correlation matrix of the network is given by:

$$Cor_{ij} = \begin{cases} \rho_D(X_i, Y_j), & \rho_D \le \rho_c \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

where $\rho_D(X_i, Y_j)$ is the distance correlation between time series X_i and Y_j.

The next step is to create the stock network. For this, we want to represent the stocks with weights that are inversely proportional to the distance correlation. Thus, we use $1 - Cor_{ij} \in [0, 1]$ as the similarity matrix that will generate the network weights.

Our strategy will use network theory to identify portfolios that have the minimal systemic risk. Thus, we define our strategy as:

Portfolio choice 1 *Given a set of assets* $\mathbf{S}$ *containing* N *stocks, we want to find weights* $\mathbf{w} \in [0, 1]^N$ *such that* w_i *is bigger if the corresponding stock has low systemic risk and* w_i *is smaller if the corresponding stock has high systemic risk.*

Our systemic risk measurement is described next.

2.2 Risk management

Our strategy aims to identify portfolios with minimal risk, so that our investment may obtain higher returns without being exposed to higher risks. Instead of using the variance as the main indicator of portfolio risk, in our strategy, we use network theory to assess the portfolio risk taking into account the systemic risk (the spread of losses due to correlation between stocks).

We build our network, as described in the previous section, with the similarity matrix defined as $1 - Cor_{ij}$. This defined the distance between the nodes in the network. Since we adopted the *Winners-Take-All*[12] method to control how our stock network is connected, we limited the node distance to a certain threshold of ρ_c. Then, we use the *Communicability Betweenness Centrality*[3] of the network to measure the portfolio risk. This measure is defined for each node as a fraction between the walks passing through this node and all the possible paths in the network. The more connections the node has the higher its communicability betweenness score. Also, the higher the score, the more the node is capable of spreading its impacts. Formally, following [3]:

Communicability Betweenness Centrality 1 *Let* $G = (V, E)$ *be a graph with* n *nodes and* m *edges with adjancency matrix* A. *Define* $G(r) = (V, E')$ *the graph constructed by removing all edges that connect to the node* r, *keeping the node itself. Let* G_{prq} *represent* *the number of paths between nodes p and q that passes through the node r and* G_{pq} *be all possible paths between p and q. Then, the communicability betweenness centrality is defined by:*

$$\omega_r = \frac{1}{(n-1)(n-2)} \sum_p \sum_q \frac{G_{prq}}{G_{pq}} \tag{2.8}$$

In our strategy, we want to allocate our capital inversely proportional to the risk of the stock.

Chapter 3

Strategy implementation

3.1 Data preprocessing

3.1.1 Acquisition of the data

The stock market is regulated by CVM but is operated by the stock exchange *B3*. The historical data is publicly available at the *B3* website[1].

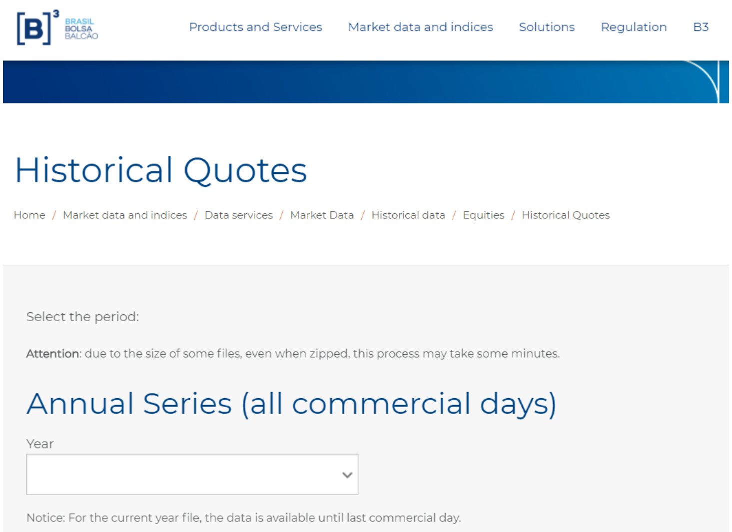

Figure 3.1: B3 publicly available historical quotes

We collected historical data from 2013 to May 2020. The data contain the trading Day, Open, Close, Low, and High prices and the Volume traded for each trading day.

[1] Available at: http://www.b3.com.br/en_us/market-data-and-indices/data-services/market-data/historical-data/equities/historical-quotes/

Table 3.1: B3 Historical data

Day	Ticker	Open	Low	High	Close	Volume	Company Name
2013-01-02	ABCB4	14.00	14.00	14.27	14.15	5 million	ABC BRASIL
2013-01-02	ALPA4	15.10	14.98	15.30	15.16	2 million	ALPARGATAS
2013-01-02	AMAR3	32.55	32.54	33.01	32.63	7 million	LOJAS MARISA
2013-01-02	BBAS3	26.00	25.46	26.19	25.80	220 million	BRASIL
2013-01-02	BBDC3	34.30	34.30	35.43	35.11	39 million	BRADESCO

The collected data for the prices time series contains six columns: Day represents the time date of the series, Open, Close, High, and Low are, respectively, the prices when the trading start, ends, the highest and lowest prices in the day and Volume corresponds to the traded amount in the day.

3.1.2 Pre-selection of stocks

We collected the historical data for all 330 companies, which may have more than one listing (for example, *Banco Itaú* have the preferred - ITUB3 and ordinary - ITUB4 stocks listed). So, we defined the following criteria to pre-select the stocks that would be added to our analysis:

1. Select only stocks that have minimum liquidity. We want our strategy to be freely available to trade the stocks on the portfolio, without incurring in liquidity risk. In this direction, we defined a threshold of average BRL 5.000,00 volume. We exclude most of the stocks in this selection, with 124 remaining.

2. Select only one type of stock per company. This allows us to remove highly correlated stocks because they are from the same company. With exclude 5 additional stocks with this criteria: *BBDC3*, *CMIG3*, *ITUB3*, *LAME3*, and *PETR3*.

3. We remove stocks with too many missing values.

4. We remove stocks that aren't connected (using $\rho_c = 0.4$) in the network.

For the last part, we did a visual analysis of the networks generated by the stocks time series and the correlation threshold. A threshold of $\rho_c = 0.15$ keeps most of the network connections alive, so we could not identify the differences between the stocks 3.2

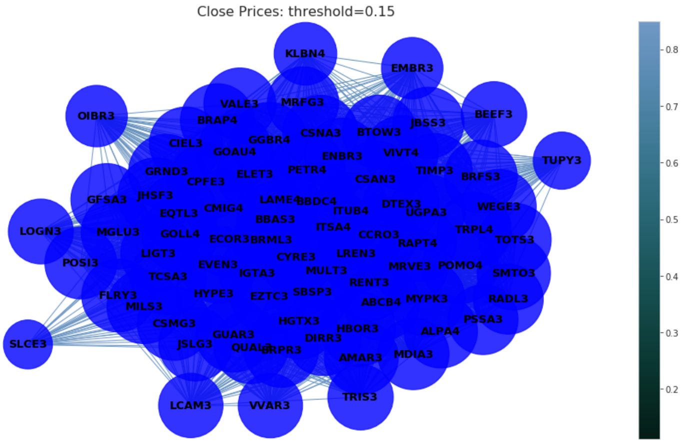

Figure 3.2: Close prices network for $\rho_c = 0.15$

For $\rho_c = 0.325$, the network still keeps too many connections, as seen on 3.3.

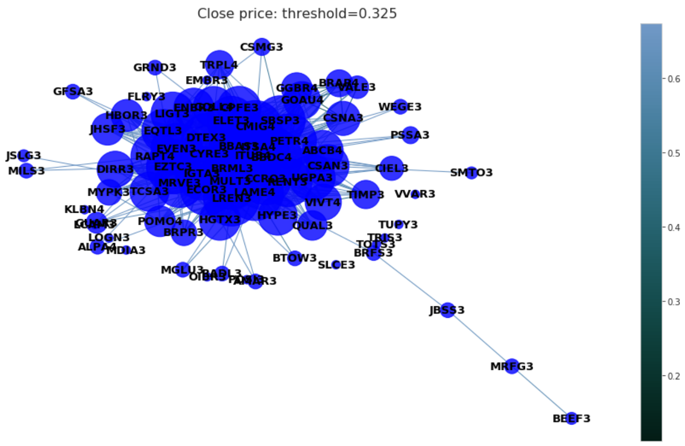

Figure 3.3: Close prices network for $\rho_c = 0.325$

Finally, for $\rho = 0.4$ we observe that we remove some of the connections and can observe different relations between the stocks. The network can be seen on 3.4.

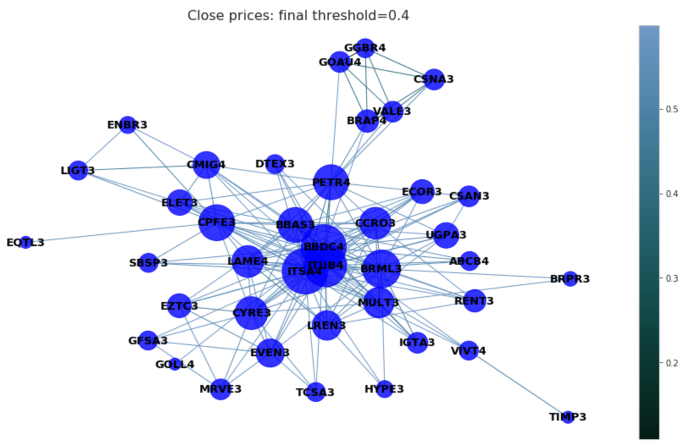

Figure 3.4: Close prices network for $\rho_c = 0.4$

After we applied the above criteria to our stocks, we ended up it a total of 40 assets that will possibly be part of our portfolio.

3.1.3 Data transformation

As described above, we use the time series of the 40 stocks to build a distance correlation matrix. We used the *Python* package *dcor*[2], and obtained the distance matrix between all stocks:

Table 3.2: Distance correlation between 5 stocks

	ABCB4	BBAS3	BBDC4	BRAP4	BRML3
ABCB4	1.00	0.20	0.16	0.44	0.48
ALPA4	0.20	1.00	0.15	0.23	0.27
AMAR3	0.16	0.15	1.00	0.24	0.27
BBAS3	0.44	0.23	0.24	1.00	0.68
BBDC4	0.48	0.27	0.27	0.68	1.00

The total distance matrix is of dimension 40×40.

Then we build the stock network using the threshold of $\rho_c = 0.4$. As seen on 3.5, the stocks on this network has an average degree (number of connections) of 9, and its distribution is skewed right.

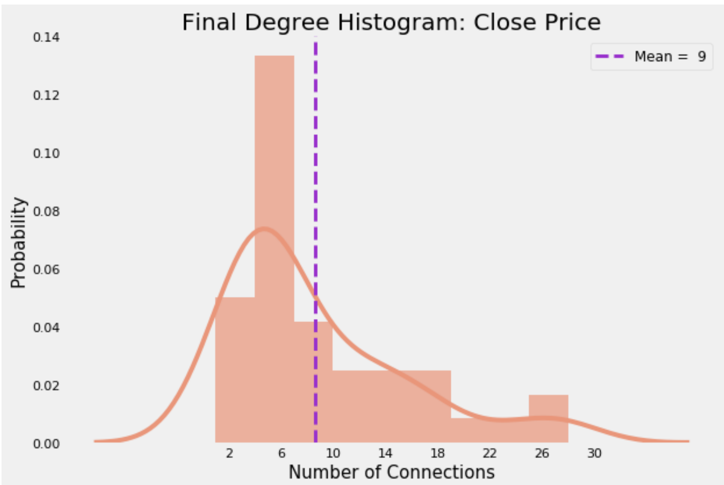

Figure 3.5: Network degree distribution

With the network built, we can start working on our strategy.

3.2 Strategy design

3.2.1 Minimal Risk Portfolio

Similar to the theory of portfolio optimization, we are interested in finding a portfolio with minimum risk. we will call this strategy the Minimal Risk Portfolio (MRP). To this

[2] https://dcor.readthedocs.io/en/latest/index.html

strategy, we start by computing the intra-portfolio risk, so that we can make investment decisions that invest less capital in riskier stocks. To compute the portfolio risk, as mentioned before, we use the *communicability betweenness centrality* measure. We can observe the portfolio risk at 3.6.

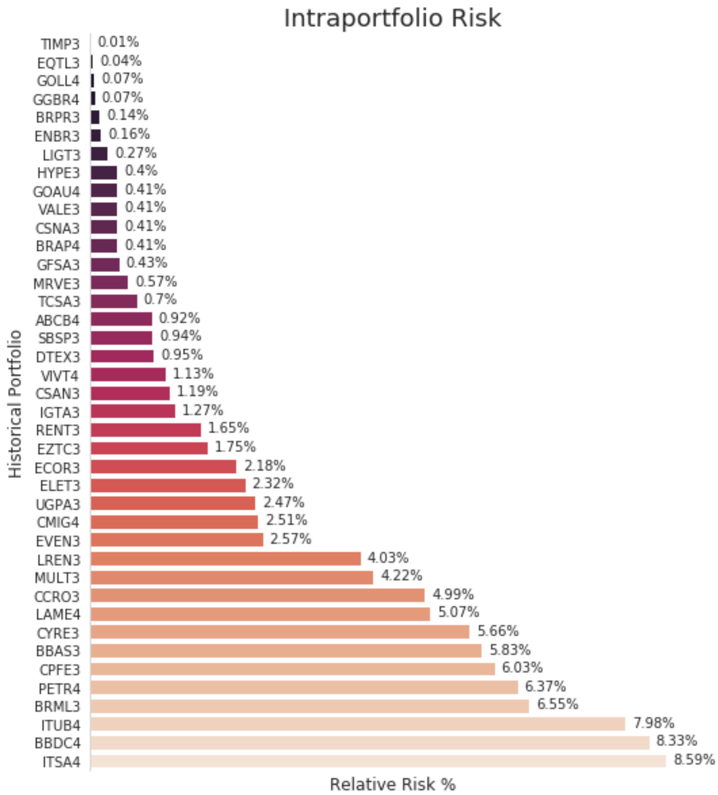

Figure 3.6: Intra-portfolio risk

We read an intra-portfolio risk plot like this: *VALE3* (*Companhia Vale do Rio Doce*) is $\frac{0.41}{0.07} = 5.86$ times riskier than *GGBR4* (*Gerdau*), *BBDC4* (*Banco Bradesco*) is $\frac{8.33}{1.65} = 5.05$ times riskier than *RENT3* (*Localiza*), ... , and *ITSA4* (*Itaú S.A.*) is $\frac{8.59}{0.16} = 53.69$ times riskier than *EMBR3* (*Embraer*)!

With this strategy, stocks that are more connected to others (more central in the netork) have the highest susception to impacts. Thus, we will invest the capital based on the inverse of the risk. We will also use a *softmax* function to smoothen the distribution and avoid investing too big a share in the least risky stock. We used a *temperature* value of 1.5:

$$\text{w'}_r = \frac{1}{\omega_r \sum_{r'} \omega_{r'}^{-1}} \tag{3.1}$$

$$e_r = e^{\frac{ln(\text{w'}_r)}{\text{temp}}} \tag{3.2}$$

$$\text{w}_r = \frac{e_r}{\sum_{r'} e_{r'}} \tag{3.3}$$

And we obtain the following distribution for a USD 10,000.00 initial capital investment:

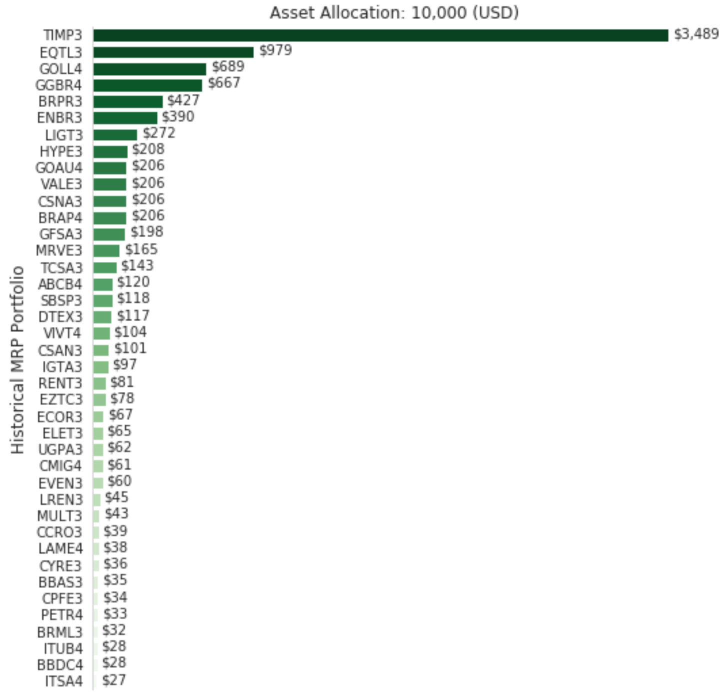

Figure 3.7: Portfolio allocation based on intra-portfolio risk

We observe that with this strategy, around 35% of the initial capital will be invested in one stock. This might seem counterintuitive since the idea is to diversify the portfolio, but according to the risk analysis, this stock is the least prone to financial impacts based on the training data.

3.2.2 Maximal Independent Set

Another strategy based on network analysis is the Maximal Independent Set (MIS)[3]. This strategy selects the non-adjacent stocks that are the most representative in the network, in such a way that the network remains connected by the selected stocks (they form a dominating set[4]).

Since the number of independent sets can be very large, instead of finding all the independent sets in order to find the biggest one (the MIS), we simulate 500 randomly selected independent sets, and from this sample, we select the maximum one.

3.3 Backtesting

In this section, we will execute our strategy on historical data. We divided the data into training (used for fitting the network) and validation (used for backtesting). The training data contains the time series from 2013-01-02 to 2016-12-29, and the validation set contains data from 2017-01-02 to 2020-05-29.

Based on the portfolio risk fitted on the training set, we assume that we make our investment on the last day of the training data, and compare our strategy with traditional ones. We will compare the results of the following strategies:

[3] https://en.wikipedia.org/wiki/Maximal_independent_set
[4] https://en.wikipedia.org/wiki/Dominating_set

- Minimal risk portfolio (MRP) (seen above)
- Maximal Independent Set (MIS)
- Efficient Frontier as proposed by Markowitz [8]

We will also compare the returns of these strategies with the historical returns of two indexes from the Brazilian market:

- *Ibovespa*: The benchmark index for the Brazilian market, representing the biggest companies listed on *B3* (currently contains 77 companies);
- *SMLL index*: Index containg the smaller companies (small cap) (currently contains 90 companies)

We have continuous allocation shares for the strategies, and we will use a discrete allocation methodology contained in the *Python* package *pypfopt*[5]. The MRP obtain the following shares distribution:

Table 3.3: MRP initial shares

Stock	Shares
BRPR3	36
ENBR3	17
EQTL3	17
GGBR4	49
GOLL4	119
LIGT3	8
TIMP3	803

By allocating multiples of the shares, we obtained the distribution shown above for the MRP portfolio. The total invested capital was $8,927.95$.

The MIS strategy will invest as below:

Table 3.4: MIS initial shares

Stock	Shares
BRPR3	40
ENBR3	20
EQTL3	19
GOLL4	136
TIMP3	909

By allocating multiples of the shares, we obtained the distribution shown above for the MIS portfolio. The total invested capital was $9,347.39$.

And the EF this one:

[5] https://github.com/robertmartin8/PyPortfolioOpt

Table 3.5: MIS initial shares

Stock	Shares
ENBR3	122
EQTL3	3
ITSA4	21
LAME4	94
TCSA3	1855
TIMP3	264
UGPA3	5

By allocating multiples of the shares, we obtained the distribution shown above for the EF portfolio. The total invested capital was 9, 986.05.

By investing in these portfolios, we can measure the performance in the validation period. The return evolution can be seen on 3.8.

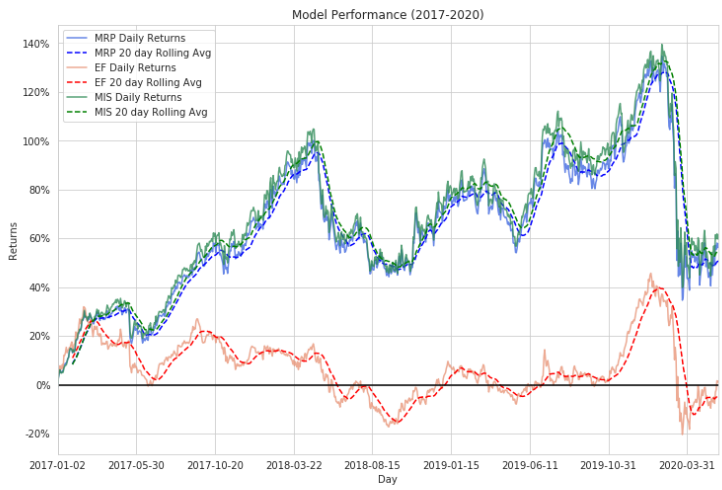

Figure 3.8: Returns on validation set

Pictured above are the daily returns for MIS (solid green curve), MRP (solid blue curve), and the Efficient Frontier (solid red curve) portfolios from January 2017 to May 2020. The color-coded dashed curves represent the 20-day rolling averages of the respective portfolios.

We can observe the following:

1. Efficient frontier has a much lower performance in the period;

2. MRP and MIS portfolios have similar dynamics;

3. The results from MRP and MIS suggests that these approaches have good performs; and

4. We want to remark that the Brazilian market was not too stable in recent years due to uncertainties in the political, economic, and social perspectives.

Next, let's observe the annual returns for each portfolio and compare them with the market.

Table 3.6: Comparison of the strategies and indexes returns on the validation period

	MRP	MIS	EF	Ibovespa	SMLL	MRP Rates	MIS Rates	EF Rates
2017	65.0%	69.5%	14.1%	26.86%	49.35%	65.0%	69.5%	14.1%
2018	70.0%	73.6%	2.3%	15.09%	8.13%	5.0%	4.1%	-11.8%
2019	117.5%	121.5%	31.7%	31.58%	58.20%	47.5%	47.9%	29.4%
2020	57.8%	61.5%	1.4%	-30.39%	-34.07%	-59.7%	-60.0%	-30.3%

MRP and MIS substantially outperformed both the Ibovespa and SMLL indexes, as well as the Efficient Frontier. The higher returns, in theory, should be obtained with the trade-off of increasing the risk of the portfolio. But as we will see in the next section, this did not occur with our strategies.

3.4 Visualizing drawdowns

Illustrated on 3.9 is the daily rolling 252-day drawdown for MIS (green), MRP (blue), and the Efficient Frontier (salmon) along with the respective rolling maximum drawdowns (solid curves).

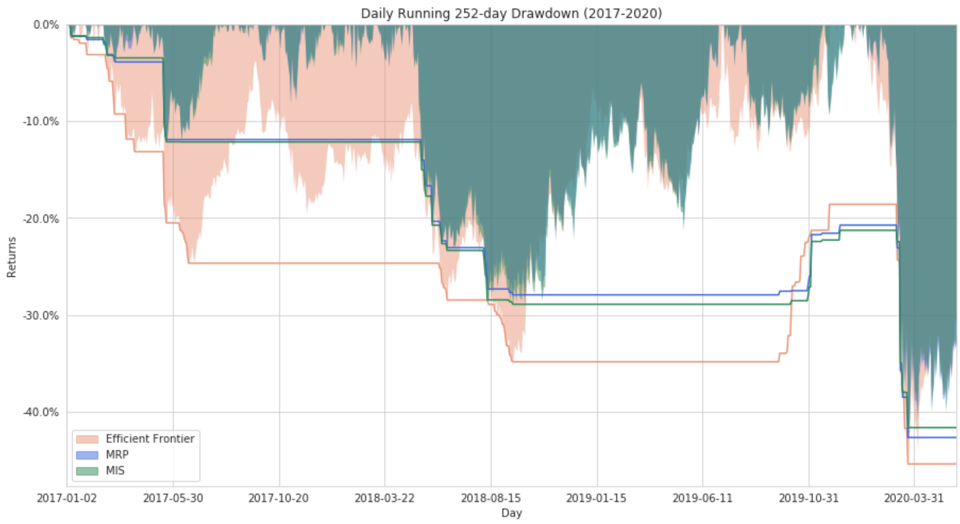

Figure 3.9: Returns on validation set

From this image, we note:

1. the MRP and MIS portfolios have significantly smaller drawdowns than the Efficient Frontier portfolio;

2. All portfolios have roughly the same maximum drawdown (around 40-45%) achieved in the *COVID-19* crisis period; and

3. MRP rolling maximum drawdowns are, on average, less pronounced than MIS. These results suggest the communicability betweenness centrality has predictive power as a

measure of relative or intra-portfolio risk, and more generally, that network-based portfolio construction is a promising alternative to the more radiational approaches like MPT.

Finally, we can compare the performance metrics for the three portfolios:

Table 3.7: Performance metrics for the three strategies

	MRP	MIS	EF
Avg Annual Rate of Returns	0.74%	1.08%	-2.39%
Annual Volatility	30.82%	31.5%	33.5%
Maximum Drawdown	-42.64%	-41.62%	-45.38%
Annualized Sharpe Ratio	-0.04	-0.03	-0.13
Returns Over Maximum Drawdown	1.36	1.48	0.03
Growth-Risk Ratio	0.04	0.05	-0.09

MRP and MIS outperformed the Efficient Frontier on every metric. These results suggest that our strategy has the potential to be used in a real-world investment.

Chapter 4

Conclusion

4.1 Conclusion remarks

We designed an algorithm to generate minimum risk portfolio (MRP) asset weights using tools from network science. First, an asset-related statistic is established, and then an appropriate centrality measure is used to extract the asset weight. As an intermediate step, we interpret the centrality score as a measure of relative risk because it captures asset volatility and their impact on the other assets in the network.

In addition, we designed a second strategy by allocating the capital by Maximal Independent Set (MIS). This strategy finds the subset of stocks that guarantee all the stocks in the network are connected. They are the most representative stocks.

Our strategies were compared with the Modern Portfolio Theory portfolio given by the Efficient Frontier (EF) method.

The portfolios were being assessed by cumulative return, rate of return, volatility, maximum withdrawal, risk-adjusted return and risk-adjusted-performance. In all performance indicators, Hedgecra algorithm is significantly better than the Efficient frontier of the portfolio and market.

4.2 Future work

Our model estimates parameters of the network based on the historical data of the training set. The use of the network betweenness centrality measure proved to be effective for minimizing the portfolio risk. Unfortunately, these dependency relations between the assets are not constant in time (the series are not stationary)[6, 5] and our model fails to adapt dynamically for different periods, especially for crisis-non crisis periods when the assets' correlations can vary significantly.

To extend our model, we can make use of advanced random processes techniques such as Bayesian sampler (for example, the *No-U-Turn Sampler*[4] or Sequential Monte Carlo [2]) to model our strategies parameters. These techniques can be extended to estimate time-varying parameters that could further improve the performance of the model in different periods.

Another approach would be to implement Copula[1] in our network, so we could model the dependency between the stock as a non-linear function that changes over time. Some

[1] https://en.wikipedia.org/wiki/Copula_(probability_theory)

new researches show results in this direction, for example, [1, 7, 9].

Finally, because the distance correlation can be applied to time series of different lengths, it would be a good choice for online applications where each stock series may have different lengths of their historical data, as well as more recently listed companies would also be ready to be added to the analysis. We want to extend our strategy to dynamically re-evaluate the portfolio periodically by updating the network when new information is available and possible change the portfolio weights and composition through time.

References

[1] Sotirios P. Chatzis and Yiannis Demiris. The copula echo state network. *Pattern Recognition*, 2012. ISSN 00313203. doi: 10.1016/j.patcog.2011.06.022.

[2] Arnaud Doucet and A M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 2009.

[3] Ernesto Estrada, Desmond J. Higham, and Naomichi Hatano. Communicability betweenness in complex networks. *Physica A: Statistical Mechanics and its Applications*, 2009. ISSN 03784371. doi: 10.1016/j.physa.2008.11.011.

[4] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, 2011.

[5] Cars H. Hommes. Modeling the stylized facts in finance through simple nonlinear adaptive systems. *Proceedings of the National Academy of Sciences of the United States of America*, 2002. ISSN 00278424. doi: 10.1073/pnas.082080399.

[6] Dror Y. Kenett, Matthias Raddant, Thomas Lux, and Eshel Ben-Jacob. Evolvement of uniformity and volatility in the stressed global financial village. *PLoS ONE*, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0031144.

[7] Dimitris Kenourgios, Aristeidis Samitas, and Nikos Paltalidis. Financial crises and stock market contagion in a multivariate time-varying asymmetric framework. *Journal of International Financial Markets, Institutions and Money*, 2011. ISSN 10424431. doi: 10.1016/j.intfin.2010.08.005.

[8] Harry Markowitz. Portfolio Selection. *The Journal of Finance*, 1952. ISSN 15406261. doi: 10.1111/j.1540-6261.1952.tb01525.x.

[9] Dong Hwan Oh and Andrew J. Patton. Time-Varying Systemic Risk: Evidence From a Dynamic Copula Model of CDS Spreads. *Journal of Business and Economic Statistics*, 2018. ISSN 15372707. doi: 10.1080/07350015.2016.1177535.

[10] Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 2009. ISSN 19326157. doi: 10.1214/09-AOAS312.

[11] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 2007. ISSN 00905364. doi: 10.1214/009053607000000505.

[12] Chi K. Tse, Jing Liu, and Francis C.M. Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 2010. ISSN 09275398. doi: 10.1016/j.jempfin.2010.04.008.