Do financial networks improve the explanatory power of the Fama-French factors? – A comparison of propagation algorithms on stock market returns

Dan Koskenniemi

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.
Espoo 14.02.2018

Thesis supervisor and advisor:

Prof. Alexander Jung
The purpose of this thesis is to investigate if modifications to the Small-Minus-Big and High-Minus-Low factors in the Fama-French three-factor model (FF3) will improve the models explanatory power in the North American stock market for the period 2000-2015, utilizing data from Wharton Research Data Centre (WRDS). The modifications of the factors are carried out by a reconstruction of the six portfolios of stocks in the SMB (Small-Minus-Big) and HML (High-Minus-Low) factors, utilizing a combination of stock correlation networks and label propagation algorithms. The theory part of the thesis is split into two major topics, the rationale and methods behind the applied investment theory and the problems that lead to further development of the Fama French three factor model, and a semi-supervised learning approach to the reconstruction of the underlying six portfolios in the Fama French three factor model. The empirical tests show that the standard Fama French three factor model does not hold a high level of explanatory power, and the graph-based semi-supervised learning approach was only able to outperform the standard FF3 model, when basic statistical tests pruned out stocks not being coherent with the theoretical description of the FF3 model.

Keywords: factor reconstruction, Fama French 3 factor model, graph signal recovery, graph-based semi-supervised learning, HML, propagation algorithms, SMB, stock correlation network
Preface

I would like to thank Professor Alexander Jung for his support and advice during the whole thesis process. Prof. Jung was the one introducing me to the graph signal recovery algorithms, which sparked my interest for writing the thesis revolving around graph-based semi-supervised learning. On a personal level, he is one of the people who really enjoys what he is doing, which has also given me the enthusiasm and drive to create a thesis I am proud of. I would also like to thank Professor Mikko Kivelä, who has taken the time to walk me through the basic concepts of networks, and introduced me to the fundamentals revolving around networks and network analysis. From the finance department at Aalto University, I would like to thank Ph.D. researcher Mikko Niemenmaa, who gave me access to the financial databases, helped me grasp the content and introduced me to the Fama French three factor model. The combination of Prof. Jung introducing the algorithms to me and Ph.D. researcher Niemenmaa’s proposing the target model (FF3) for my research, later formed the foundation of the whole thesis. From the Big Data group, I would like to thank Master of Science student Alexandru Mara, who has both helped me with explaining the functionality of the algorithms, and became a closer friend of mine during the writing process. Finally, I would like to thank my family for all support and my friends for being great listeners to my talks and ideas (sometimes too many) about the thesis and the work revolving around it. I wish all the best to people that helped and supported me during the process. Thank you!

Otaniemi, 14.02.2018

Dan E. Koskenniemi
## Contents

Abstract ii  
Preface iii  
Contents iv  
Symbols and abbreviations v  
1 Introduction 1  

2 Financial Theory, Asset Pricing and the Fama French Three Factor Model 4  
2.1 The Notion of Value and Price 4  
2.2 Investment Frameworks 10  
2.3 Investment Models 13  

3 Reconstructing the Underlying Portfolios using Graph-Based Semi-Supervised Learning 19  
3.1 Graph Based Semi-Supervised Learning 20  
3.2 Label Selection 21  
3.3 Networks 23  
3.4 Graph Signal Recovery 29  

4 Research Methods 33  

5 Results 38  
5.1 Stock Correlation Network and the Six Underlying Portfolios 38  
5.2 The SMB- and HML-factors 46  
5.3 Explanatory Power of Regression Models 49  

6 Summary and Conclusions 55  

References 57  

A Additional Tables 61
Symbols and abbreviations

Symbols

\( \alpha \) Excess return
\( \beta \) Variable of linear regressor
\( \phi \) Variable of multiple linear regressor in the FF3 model
\( \rho_{xy} \) Pearson correlation coefficient-value between two arbitrary stock returns
\( \sigma \) Standard deviation (or volatility)
\( \sin(\theta) \) Absolute correlation distance
\( E[P_t] \) Stock price at time \( t \)
\( E[R_a] \) Expected return of security \( a \)
\( N \) Number of nodes
\( R_F \) Risk-free rate
\( R_M \) Market Return
\( r_t \) Stock return at time \( t \)
\( R_P \) Portfolio return
\( t \) Time
\( W_{ij} \) Correlation matrix

Abbreviations

BP Big portfolio
BV Book value
CAPM Capital Asset Pricing Model
CML Capital market line
EMH Efficient market hypothesis
EWR Equal weighted return
FF3 Fama French three factor
GB SSL Graph-based semi-supervised learning
HML High-Minus-Low
IPO Initial public offering
LPA Label propagation algorithm
ME Market value of equity
MPT Modern portfolio theory
MSE Mean square error
MST Minimum spanning tree
PCC Pearson correlation coefficient
SLPA Sparse label propagation algorithm
SMB Small-Minus-Big
SP Small portfolio
TA Total assets
TL Total liabilities
WAMC Weighted average market capitalization
WRDS Wharton Research Data Centre
1 Introduction

The stock market has long been an intriguing challenge for economists, physicists and mathematicians trying to forecast the performance of stocks. As it is one of the most lucrative and brutal industries on earth, people with extraordinary scientific contributions have taken on the challenge. For instance, Sir Isaac Newton, the inventor of calculus and three laws of motion, did not perform well on the stock market and lost millions. When such a heavyweight scientist fails to explain stock returns, predicting returns in the stock market is clearly a very tough challenge.

Fast-forwarding to the 20th century, theories and frameworks, such as the efficient market hypothesis and modern portfolio theory, have lead to descriptive models such as the capital asset pricing model and the Fama French three factor model, in attempts to explain stock returns. To date, there still exists a vigorous debate regarding the assumptions and correctness of the frameworks and investment models, due to unexplainable factors or anomalies. In this debate, one side argued that markets are rational and quantifiable, whereas the other side argued that investment decisions are based on other factors, which are not quantifiable, such as human psyche. As a consequence, subfields within investment theory started to emerge during the of the 20th century, where two of the largest ones are econophysics and behavioral finance. Due to large increases in computing power and the expansion of advanced data analysis techniques and artificial intelligence, the possibilities for investigating and testing these theories are multiple times higher than during the lifetime of Sir Isaac Newton. By incorporating network theory into data analytics, construction of stock correlation networks has had successes in disproving or validating investment models[6] or identifying stocks dominating the movement of other stocks on the market[28]. Due to earlier successes in capturing information from market dynamics on the stock market, I will investigate whether a combination of a stock correlation network and graph signal recovery algorithms will improve the explanatory power of the Fama French three factor model[14] (FF3).

The FF3 theory has proven to have a certain explanatory power of stock returns[14]. To improve the model further, I will challenge the standard construction of the factors in the Fama-French model and redesign the content of the six underlying portfolios, using graph-based semi-supervised learning. In detail, the graph-based semi-supervised learning approach will utilize a combination of: 1) initial stock samples from the standard FF3 model, 2) a stock correlation network and 3) graph signal recovery algorithms, to redesign the content of the portfolios. To the best of my knowledge, this is the first-time graph-based semi-supervised learning has been used to modify an existing model striving to explain stock returns.

The graph-based semi-supervised learning approach will construct a network of the stocks based on the correlations between their respective returns. The hypothesis I present is that such networks will create more coherent underlying portfolios, than the standard way of constructing the portfolios in the FF3 model. Based on this hypothesis, portfolios created by graph-based semi-supervised learning should improve the explanatory power of the FF3 model, since the created stock correlation networks should create portfolios including stocks exhibiting more similarities in
their performance characteristics than stocks based on the standard model. The argument supporting this assumption is based on two anomalies behind the FF3 factors: the size effect and the value effect, which have historically measured certain diverging behavior for certain kinds of stocks. For instance, Fama and French[14] showed that there exists a size-anomaly in the capital asset pricing model, where companies of the same size had high correlation in their stock returns. The same can be said for the value effect, where companies with close Book-to-Market ratios had a high correlation in their respective stock returns.

The research question is whether the hypothesis is true, in other words, whether underlying portfolios created with graph-based semi-supervised learning explain returns on the market better than the standard model. If the hypothesis is true, one or both of the graph signal recovery methods redefining the underlying six portfolios in the FF3 model will provide a higher statistical explanatory power of stock returns than the standard model.

The hypothesis will be tested through empirical and quantitative methods on monthly returns of stocks listed on the US stock market for the time period 2000-2015. The research part was executed using Matlab for analytics and MySQL for data storage, and the data utilized was retrieved from Wharton Research Data Centre[58]. The standard model, i.e. the method for constructing the factors in the FF3 model, was created using the same data, methods and restrictions described in the Fama French article[14]. In the construction of the network models, two hyperparameters were utilized: sample size and correlation length. Sample size described the percentage of stocks that should be picked from the portfolios in the standard model, and correlation length described length of return history that should be accounted for when constructing the stock correlation network. The stock correlation network was set to keep edges with high, positive correlation values, to minimize the amount of noise in the network for graph signal recovery. After the three model categories had constructed the underlying portfolios and assembled the FF3 models: 1) standard FF3, 2) graph signal recovery method: label propagation FF3 and 3) graph signal recovery method: sparse label propagation FF3, the returns from stocks on the market were fit with multilinear regression (as described in the FF3 article[14]) against the models. Thereafter, the adjusted R2-value averaged over all stocks was used as measurement for explanatory power. The presumption is that the FF3 model is correct as is, and the only improvement induced is the set of stocks in the underlying portfolios. Therefore, the experiments will be conducted firstly under this presumption, and if there are cases where the presumption does not hold, the experiment will be repeated by pruning such stocks from the analysis, without further inspection regarding the reasons why the model may or may not fail.

This thesis consists of six parts. The first part introduces the topic, the study and describes the structure of the thesis. The second part presents the financial theory behind the Fama-French three factor model and asset pricing in general. The third part describes how the underlying portfolios were reconstructed using a graph-based semi-supervised learning framework. The fourth part describes how the empirical and quantitative research methods were applied, including how and what data was collected, how such data was used, and describes the choices of measurements and
parameters used. The fifth part presents the results of the regressions for each of the three model, the differences for the SMB- and HML-factors for each model and the results of the comparison made between the three models and, finally, the sixth part discusses the study and its results.
2 Financial Theory, Asset Pricing and the Fama French Three Factor Model

This chapter discusses how price and value is formed, how theoretical frameworks have emerged to explain the market as an entity, how movement of individual stocks can be explained by factors affecting the market, and, finally, the rise of the Fama French 3 factor model as a result from critique against previous frameworks and models.

2.1 The Notion of Value and Price

The history of finance contains several different approaches to the analysis of asset prices, financial markets, behavioral aspects and the like in building an understanding of: 1) how the markets for financial assets function and 2) what the drivers behind market dynamics are. According to Gurley & Shaw[24], the logical way for an economist to study finance is to study it as a market problem. What the market problem means, is that the analysis should start by finding factors that determine the demand and supply for any financial asset, including an understanding of a market equilibrium. Whereas such equilibrium is slightly out of the scope of this thesis, it is important to understand the importance the equilibrium state plays in the context of theoretical frameworks. Each set of demand, supply, and market-equilibrium equations defines a market, which may be analyzed as a whole entity or divided into subparts. The combination of all such equations for all assets and markets sets the basic ground for finance.[24]

Markets for Financial Assets

The financial market serves as a market, where people with excess funds can transfer funds to people with a shortage of funds. A well-functioning financial market has a direct impact on personal wealth, the behavior of businesses and even the cyclical performance of the economy.[38] To further simplify why financial markets matter, let’s assume that you have come up with an invention that could save time and effort for local farmers. As an inventor without lots of excess capital, you realize that your invention could be built at a prototype stage. However, you also realize that you won’t be able to finish your product and develop a fully functioning business without further capital. The shortage of capital would lead to the fact that your dream of taking the invention to the market would stop there, and your efforts would be in vain. Now, let’s look at the other side of the spectrum, where we have a well-funded millionaire with a lot of excess funds. When these funds are not invested, they would yield zero returns for the millionaire, who might have a higher risk appetite than simply sitting on his funds. The probability of the millionaire knowing you would also be very slim, which is why 1) you would not be able to raise funds from the millionaire and bring an invention to the market and 2) the millionaire would not be able put his funds to efficient use. As can be seen, there is a problem that needs to
be solved here, demonstrating why a financial market is crucial; excess capital can be connected to projects or companies that are expected to yield a positive return.

The form, method and object of the capital investment is always subject to several factors, such as the 1) risk appetite of the investor, 2) the time horizon for such investor’s investment, 3) regulatory considerations. When both you and the millionaire have agreed on terms for the millionaire lending you the funds, the supply and demand side of the financial market meet in an equilibrium, as discussed above. This example highlights the importance of the financial market. The efficient allocation of capital contributes both to higher production as well as access to capital and, consequently, also improves the overall efficiency of other markets than the financial market.[37] In the example above, there needs to be a contract to satisfy both your and the millionaires needs, around which the transaction will be built. The millionaire needs to be sure that his investment is protected and that repayment is secured, whereas the inventor needs to have protection from further demands on lend capital. Otherwise, the market would only be built on trust and stay very local. To ensure that both parties’ needs are met, there needs to be both a contract and proof of the transaction, holding monetary value; they are called securities. In this thesis, the focus will be on stocks, which is a corporate security.

A stock (or a share) is a security which is part of the capital structure of a company, providing the holder of such security with one or more of two sets of rights, administrative and financial rights. The administrative rights contain the general voting right a holder of one stock has in the decision making of the company (as part of the shareholders’ meeting). The financial rights allow, depending on the specific terms for such stock, the holder to receive a portion of the company’s annual result and to claim a share of all the company’s (net) assets. In other words, the holder of a stock is an owner of the company. There exists other securities in a company’s capital structure, such as debt and hybrid instruments. Whereas especially debt instruments do not grant any ownership rights in a company, they give their holders a higher priority claim on the assets of the company, if the company becomes insolvent. These claims, therefore, hold a higher priority to be paid, than the claims of the stockholders.

When comparing the two securities, obligations or debt securities are, therefore, usually considered less risky due to their higher ranking priority to be paid back. The concept of risk is an important distinguisher between the categories of securities being issued by a company, since it is directly connected with the financing of the company’s operations. The securities are connected to the financial markets by firstly being issued at a price in the primary market, and, secondly, by being sold for a price (between two investors) in the secondary market.[38]

Valuing a Company

Earlier we have discussed what financial markets are, why they are important for the economic system and how the transaction of funds is carried out. The next topic is about how the value of the stock is formed and the logic behind this process. To understand how value is formed, let’s start with an example where you now have
developed a fully functioning business. The business is going well and you are about to announce a second model of your invention, but you are again finding yourself facing a shortage of capital. To finance your model, you have decided that you need to sell a part of your stocks in the company. This action is necessary to cover the expenses of research and development of the second model, as well as launching the product to the market. In the process of launching your first batch of stocks to the market, you need to know how much your company is worth. Since the central focus in this thesis will be on stocks, the discussed mechanisms of company valuation will primarily focus on how the value of the stocks is determined. Value and price are obviously two different concepts, since value can, in theory be established as a data point, whereas the price is a function of the market and where the equilibrium lands, as described in the previous chapter. The process of valuing a company and its different parts supports both the company’s management’s decision making as well as the analysis and price establishment by the investor. According to Fernandez[20], the methods for valuing companies can be classified in six groups:

Table 1: Company valuation methods.

<table>
<thead>
<tr>
<th>Balance Sheet</th>
<th>Income Statement</th>
<th>Mixed Goodwill</th>
<th>Cash Flow Discounting</th>
<th>Value Creation</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book value</td>
<td>Multiples</td>
<td>Classic</td>
<td>Equity cash flow</td>
<td>EVA</td>
<td>Black&amp;Scholes</td>
</tr>
<tr>
<td>Adjusted book value</td>
<td>PER</td>
<td>Union of European Accounting</td>
<td>Dividends</td>
<td>Economic profit</td>
<td>Investment option</td>
</tr>
<tr>
<td>Liquidation value</td>
<td>Sales</td>
<td>Experts</td>
<td>Free cash flow</td>
<td>Cash value added</td>
<td>Expand the project</td>
</tr>
<tr>
<td>Substantial value</td>
<td>P/E,EBITDA</td>
<td>Abbreviated income</td>
<td>Capital cash flow</td>
<td>CFROI</td>
<td>Delay the investment</td>
</tr>
<tr>
<td></td>
<td>Other multiples</td>
<td>Others</td>
<td>APV</td>
<td></td>
<td>Alternative uses</td>
</tr>
</tbody>
</table>

In this thesis, I will shortly explain the differences between these methods, but later I will focus mainly on the balance sheet valuation method concerning book-value, since it is one of the metrics used in the Fama French three factor model discussed later. The listed balance sheet-based methods seek to determine the company’s value by estimating the value of its assets. These methods do not take into consideration the expected future value of the company or other factors that may have an impact on the value of the company and its securities, but rather only focus on the present state of the company.

A company’s book value, or the net value of its assets, is the value of the shareholders’ equity stated in the balance sheet (capital and reserves), or the difference between total assets and liabilities. This method has some shortcomings, such as the difference between accounting valuation methods and market defined valuation criteria in the process of stating the “fair value” of an asset or a company. Consequently, these two values rarely converge.[20]

The income statement based methods again focus on the income statement of the company, rather than its balance sheet. According to these methods, the value of a company is established through the size of its revenue, earnings or some other metric. These metrics are typically also compared to those of similar companies in order to obtain a better understanding of how the company should be valued. An example of this would be to state that in a preselected industry, the value of a company should be established as a multiple of its earnings.

In the goodwill-based methods, there is an attempt to measure the value of a company’s intangible assets, which contribute to an advantage of the company
in relation to its competitors. These methods typically start by using a balance sheet-based method and add the goodwill-based value to it.

The cash flow discounting methods, which are very widely used, base their estimate of the company’s value on its future expected cash flows and the risk taken to generate such cash flows. These cash flows are discounted, accounting for time, to the present in order to get a present value of all such cash flows.

Economic value or profit methods focus on the difference between the cost of capital of the company and the profit it generates, to establish the residual value generated between profit and costs.

Finally, the options based methods try to establish the company as a set of (future) projects and produce a decision tree around whether to proceed with one or another option, such as investing in expansion of a business unit, or selling it.

Previously we have discussed the different methods through which a private company can be valued. To understand how value is formed when a company becomes public, let’s continue with the example above in a situation where you want to capitalize on your business and sell at least part of it in a bigger transaction. In the process of launching your first batch of stocks to the market, you need to state a price based on the value of your company. To objectively evaluate the value of a company, an external team of experts is hired to investigate the total assets of your company as well as projected cash flows from future operations. The process of evaluating the value and issuing stocks to the public for the first time is called an Initial Public Offering (or IPO). After the IPO, there are multiple ways of evaluating the value of a company. In this thesis, we will primarily focus on two methods: book value and market value. Book value is determined solely based on the balance sheets of the company, and is the difference between company’s total assets and total liabilities.

\[ BV_t = TA_t - TL_t \] (1)

Where \( BV \) is the book value, \( TA \) total assets and \( TL \) total liabilities of the company. If for instance your total assets would be 100 million and the liabilities 60 million, the book value of the company would compute to 40 million. After the company has gone public, it is also possible to determine the company’s Market Capitalization or Market Value of Equity, defined as:

\[ ME_t = P_t S_t \] (2)

Where \( P_t \) is the price for the stock at time \( t \), \( S_t \) is the total amount of shares at time \( t \) and \( ME_t \) is the Market Capitalization at time \( t \). The market value of the stocks will be more relevant after the IPO has been finished and the stock has had some history on the market, but notable is that both the book value and the market value are dynamic (both will vary over time). When comparing the two methods of valuing a company, there are typically two basic scenarios. If the book value is higher than the market value, it means that the market is valuing the company for less than its stated value according to the balance sheet. The previous scenario is usually an
indication that the market has lost confidence in the company being able to generate profits in the future, and hence sees the stock investment as a bad decision. A higher market to book value indicates that the balance sheet is not properly reflecting the value that the investors see in the company, such as a well-functioning company with consistently steady profits in history and great future projections. Both valuation models are varying over time, and the effect of time on stock prices, will be covered in the next chapter.

Stock Prices and Stock Returns

In the previous chapter we discussed how the value of the company is determined, how a company enters the stock market and the separate methods to evaluate company value. Let’s now assume that the IPO have went well and you have started to see that the second invention yields positive cashflows for the company, increasing the book value of the company. The response on the market is that the investors from the IPO have seen an increased value of the stock, and hence do not easily sell the stock due to the success of the second model. On the other side, more investors are gaining interest in the company due to the positive history and growth of the company, which also results in investors being ready to pay a higher price for the stock. The original investors would like to gain as much capital as possible from the batch of stock, which means that the higher the price of each stock is, the better. From the buyer’s perspective, there are a lot of investors interested in shares of your company. However, everyone still wants to get the stock as cheap as possible, since there is no guarantee that the future value of the stock will cover the initial investment on the stock. This leads to a conflict of interest between both the investors from the IPO and the potential new investors, since the investors from the IPO wants as high a price for the stock as possible, and the new investors wants to keep the price low. This issue is solved by ensuring the equilibrium state between stockholders and new potential stockholders with a stock auction.

It meets both parties’ needs, since the stock will be sold to the highest bidder and hence yield as high a stock price as possible for the current stockholders. The potential stockholder in turn receives the stock for a price which is in the boundaries of being profitable. The definition of the price of the stock on the market, is when the buy side and the sale side come to an agreement on the value of the stock, where the sale side is called supply and the buy side demand. Demand stands for the buyer’s willingness to invest in the stock, and supply is the limited amount of stock issued to the sellers.[3] When the buyers and the sellers comes to an agreement on the price of the stock, supply and demand have reached an equilibrium where the transaction can take place. This can be seen in Figure 1 below as a hypothetical price for Intel Corporation stocks:
As can be seen from Figure 1, the equilibrium price of Intel Corporation’s stock at a certain point in time is 25.0$/stock. Now, depending on the total amount of existing stocks for Intel Corporation, it is possible to derive The Market Value of Equity for Intel Corporation. The fact that the stock is worth more after the IPO, is why people are interested in investing on the stock market rather than keeping it on the shelves, since it will increase their own funds. Let’s now assume that the second version of the invention has been released to market, and the invention has been flying off the shelves resulting in an increased positive cashflow for the company. The investors have noticed the company’s increased results, and are now willing to pay a higher amount for each stock. As a result, the company is also worth a lot more due to the increased cashflow. What drives the value of a stock, depends on the market views of the stocks. It can be thought of as people’s perception of how the value is going to change in the future, which can both drive the price up or down. The change in stock price is defined as follows:

\[
\nabla P_t = P_t - P_{t-1}
\]

As stated earlier, the investors are investing in stocks, since they are expecting that the stock would increase their initial investments value. In other words, the investors want to get a return for taking risks by investing in a company’s stock. Assuming that the company does not pay any dividends and retains all earnings, the return on a stock is the percentual change between two price points as defined below:

\[
r_t = \frac{\nabla P_t}{P_{t-1}}
\]

From an investor’s point of view, it is worth noting that during a purchase of a stock, the buyer is always buying the stock at a certain point in time, and hence
needs to speculate on the stock’s movements in order to achieve positive returns. Since all stocks on the market can go either up or down, this implies that the buyer also needs to have information about how the price of the stock will develop in the future. In order to purchase a stock, the buyer-side must see a higher value of the stock in the future, than the stock reflects at point Pt (for instance the information of a second version of the invention being released to the market). These assumptions are based on the investor being completely rational in his or her decision, which has lead to the need to model the forces affecting formation of equilibrium and as a result, purchase order execution.

2.2 Investment Frameworks

Like the earlier example of you developing your investment to a fully functioning business with shares that are offered on financial markets, there are similar companies with different backgrounds that have sparked to fully evolved businesses. The aggregate of all businesses with publicly available stocks are called the stock market.[37] As time passes by, investors are looking for the investments that will yield the highest possible future returns, which is why modeling returns is a highly discussed domain in academia. This chapter discusses how price and value is reflected onto decisions made by investors on the market.

The Efficient Market Hypothesis

One of the most famous frameworks within this domain is the Efficient Market Hypothesis (EMH) presented by Eugene Fama in his doctoral dissertation. EMH states that an investor by time can not get a higher return than the total return of the market (all returns from each company put together). This is a result of EMH stating that the Market Value of Equity for the companies will be a fair one and there is no possibility of under- or overvaluation. The only way to gain a higher return than the market in total is to invest in riskier stocks. The EMH is based on the perfect market assumptions, which include the following: 1) no transactions costs, 2) information is easy to obtain and free, 3) investors have homogenous expectations, 4) investors are rational and therefore markets are efficient.[21] An efficient market is one in which prices of securities fully reflect available information. In practice, this would mean for our previous example at the time when the invention was released, that every investor would: 1) be able to purchase the stock without anyone taking a cut from the purchase order, 2) be aware of the release of a second invention, 3) have the same projections of the future cashflow that the invention would bring and 4) act according to their valuations and projections and would not favor or discard the company in any manner affecting the investment decision. Furthermore, the EMH consists of three subsets:

1. Weak form: The weak-form of EMH states that the market is efficient and that all the market information is reflected on the stock price. The weak-form also assumes that past returns compared to future returns on the market are
independent. In other words, it will yield no benefit to analyze earlier prices of stocks in order to determine the future price of the stock.

2. Semi-Strong form: The semi-strong form includes: 1) an assumption that the stocks on the market will adjust quickly to new information, and 2) all the assumptions of the weak-form EMH. Given the assumption that stock prices immediately react to new available information and investors act to purchase stocks after this information is released, an investor cannot benefit over the market by trading on new information.

3. Strong form: The strong-form EMH implies that the perfect market holds: the stock prices reflect information both public and private, and includes both the weak-form EMH and the semi-strong form EMH assumptions. Since it is assumed that stock prices reflect all information (publicly announced and inside information), it is not possible for an investor to gain more profit than the average investor.

Whereas the EMH has been actively discussed and researched in academic literature, there seems to be a strong view that the theory is not accurate.[13] One of the arguments behind this challenge is the notion that investors are rational and hence react in a similar, structured manner to new information. This argument is based on behavioral sciences and challenges the notions of fundamental analysis. In fundamental analysis, the proponents argue that share prices reflect the future expectations correctly, as such prices are updated by the reactions of rational investors to new information regarding the parameters affecting the stock price, such as interest rates, risk measures, future expectations, industry regulation and the like. The fundamental value of the stock is hence defined as the net present value of all future, expected, cash flows to investors.

The completely opposing theory to fundamental analysis, and which is strongly opposing the notion of rational actors, is the methodology analyzing the markets through psychological and similar theories, forming the part of finance theory known as behavioral finance. One of the most famous theories is Keynes’ “animal spirits”, which is a statement from his famous book The General Theory of Employment, Interest and Money[29], where he related the instability in markets not only to speculation, but also to the “characteristic of human nature”. According to these theories, and contrary to the fundamental analysis, the stock price follows a formation process not driven by established valuation techniques and the rational actions and decisions of the investors, but on the (mass) states of the human investors putting emotions and irrational biases in play, instead of “mathematical expectations”, as Keynes wrote. This also explains why speculative bubbles arise, since investors hope that they can sell tomorrow to someone else for an even higher price. In other words, the investor hopes that the speculative bubble continues to grow and bases his investment decision on such hope, rather than on proper analysis and actions.[29]

Anomalies are another challenge to the EMH and are defined as empirical results that do not correspond to the theories of asset-pricing behavior, indicating either market inefficiencies (directly opposing arguments to the EMH) or imperfections
in the pricing model used. Anomalies have been researched and documented for decades, however, they seem to exhibit a disappearing nature. The main reason for this is likely that inefficiencies indicate profit (trading opportunities) and probably disappear due to the amount of traders trading on such inefficiencies, essentially making the anomaly part of the EMH, as investors react, presumably rationally, on such information.[33]

**Modern Portfolio Theory**

The theory of Modern Portfolio Theory (MPT) is a theory developed in the 1950’s by Harry Markovitz, stating that the rationale driving an investment decision should be based on two factors: 1) return and 2) risk. Risk in this context can both be positive or negative, and is per MPT related to the variance of the returns of a stock. According to MPT, investors are making investment decisions based on maximizing future returns. On the other hand, the investor wants to minimize risk as much as possible, since it brings uncertainty of the future return of the investment.[36] When comparing a government bond to a startup company, the government bond’s expected return is easier to estimate compared to a startup company with zero market history. For an investor, it is much harder to predict how the startup company will turn out, and therefore, the return of an investment in the startup company could be 1) huge or 2) end up going to zero. Therefore, it is crucial for the investor to balance the amount of risk he/she is willing to take versus the required returns for the investor.

Even though there exists a certain amount of risk in a security, the risk and return for the combination of multiple securities (portfolio) is different. Each portfolio yields a certain return, but according to MPT, the risk is necessarily not the same for every portfolio with the same return profile. To maximize average return of the portfolio while minimizing risk it is possible to apply diversification, which is a technique selecting securities that are as much uncorrelated in their risk profile as possible to avoid "putting all the eggs in one basket". The portfolio with maximum return and minimum risk for each return profile is called "The Efficient Frontier" as can be seen below in Figure 2.
Figure 2: The Efficient Frontier and Capital Market Line according to MPT.

All the portfolios on the efficient frontier are called efficient portfolios, and if a tangency line would be drawn on the efficient frontier, it would nod one of the efficient portfolios. This line is called the Capital Market Line (CML), where inefficient portfolios would be under the CML whereas efficient portfolios would lie along the CML. The CML is defined as follows for each of the efficient portfolios:[12]

$$\bar{R}_e = R_F + \frac{\bar{R}_M - R_F}{\sigma_M} \sigma_e$$

(5)

Where $R$ denotes the return, $\bar{R}$ denotes the average return, $\sigma$ the standard deviation, subscript $e$ denotes an efficient portfolio, subscript $M$ denotes the market and $F$ the riskfree rate. The term $\frac{\bar{R}_M - R_F}{\sigma_M}$ is the market price of risk for all efficient portfolios. To qualitatively express each term:

(Expected return) = (Price of time) + (Price of risk) * (Amount of risk)  

(6)

Which leads to one of the groundbreaking formulas of its time and is widely used in industry today: Capital Asset Pricing Model or CAPM.

2.3 Investment Models

As frameworks started to emerge regarding the investment choices investors make, and drivers behind them, it lead to the development of models striving to describe and forecast the behaviour of stock returns. One of the earliest ones is the Capital Asset Pricing Model, which describes the expected return of stocks by looking at the movement of the stock market.
Capital Asset Pricing Model

The Capital Asset Pricing Model (CAPM) is a model developed by Sharpe[52] and Treynor[54], and extended and clarified by Lintner[32] and Mossin[40]. The model builds up a relationship between the security and the whole market with certain prerequisites. The prerequisites are the following: 1) investors are solely investing by maximizing return and minimizing risk, 2) there are no taxations or transaction costs on the investment, 3) investors are viewing the securities with the same probability distributions and 4) investors can borrow and lend funds at a riskless rate of interest.[25] The CAPM-model is presented below:

\[
\bar{R}_a = R_F + \beta_a (\bar{R}_M - R_F)
\]  

(7)

Where \( \bar{R} \) stands for average expected return, R for realized return, subscript a for an individual asset, subscript F for riskfree rate and subscript M for the market. The idea behind the CAPM is that investors must be compensated in two ways: 1) time value of money and 2) risk of investment. Time value of money is compensated in the form of the risk-free rate, which in practice is usually measured using a government treasury bill. The risk of investment is derived by comparing the movement of the stock versus the market, which translates to using linear regression between the asset and the market over a period of time. The model has sparked vigorous debate in academic circles regarding its explanatory power of assets. Fama and French[14] claim "CAPM is useless for precisely what it was developed to do." Based on empirical results from their paper, Berg[4] cites Fama's claim "beta as the sole variable explaining returns on stocks is dead." Fama and French[17] continue challenging the CAPM in "The CAPM is wanted, dead or alive." Fama and French[18] suggest that problems with CAPM may be a causality of the simplifying assumptions and the difficulties of estimating the market portfolio in the tests of the model. Fama and French are two of the most prominent researchers in this field criticizing the capital asset pricing model. Some of the most famous documented anomalies of the CAPM are (1) the turn-of-the-year effect, (2) the weekend effect, (3) the momentum effect, (4) the size effect and (5) the value effect. In two papers from 1983 (Reinganum[49] and Keim[27]), it was shown that much of the abnormal returns for small companies occur around year end, and especially during the first two weeks of the year (also known as the January Effect). Roll[50] argued that one reason behind this anomaly might be tax driven, so that losses would be realized prior to year-end (and the stocks bought back again in January). French[22] found a similar kind of seasonality effect in the weekend effect. Whereas the reason for the effect has been much discussed, it is speculated that (historically) there has been a drive to sell before the weekend, to not have to carry the risk in relation to the stock over the weekend, when the markets are closed. Both the turn-of-the-year effect and the weekend effect seem to have disappeared, or at least significantly diminished since originally discovered.[51]
Fama-French Three Factor Model

Eugene F. Fama and Kenneth R. French are two professors that are active within portfolio theory and asset pricing of financial securities. Their research received attention for its critique against the CAPM, wherein they questioned the ability of the CAPM to explain the average return of securities with a single variable: market risk premium. The validity of CAPM is also questioned in empirical studies carried out by Breeden et al. and Reinganum. As a competitor to the CAPM, Eugene Fama and Kenneth French developed the three-factor model to enhance explanatory power of the model regarding the value of stocks and bonds. Instead of utilizing \( \beta \) as the single explanatory variable of asset returns, Fama and French incorporated two additional risk-factors related to the size and the valuation of the company. The two added factors were Small-Minus-Big (SMB) and High-Minus-Low (HML), where SMB studies the Market Value of Equity (ME) and HML the Book-to-Market Value (BV/ME) in stocks. The aim of the model is to increase the explanatory power of returns in individual assets, by encapsulating additional risks not considered in the CAPM. The Fama and French Three Factor Model (FF3) is defined as follows:

\[
E[R_a] - R_F = \phi_1(R_M - R_F) + \phi_2(SMB) + \phi_3(HML) + \alpha + \epsilon \tag{8}
\]

Where \( E[R_a] \) stands for the expected return of a stock, \( R \) for return, subscript \( a \) for a single security, \( F \) for the risk-free rate, \( M \) for the financial market, SMB and HML for the risk-factors, alpha for excess returns and epsilon the average error in regression analysis. Since FF3 will be in focus of this paper, let’s investigate the terms further.

Market Return. The market return will be computed according to the Weighted Average Market Capitalization (WAMC) defined below:

\[
R_M = \frac{1}{ME_M} \sum_{a=1}^{n} ME_a R_a \tag{9}
\]

Where \( R \) stands for return, subscript \( M \) stands for the financial market, \( ME \) stands for Market Value of Equity, subscript \( a \) for all assets belonging to the market and \( n \) for the amount of stocks belonging to the market. The difference between Equal Weighted Return (EWR) and WAMC is that each stock has a certain weight in the portfolio for WAMC, whereas the weight is equal for all stocks using EWR.

Risk-free Rate. The risk-free rate stands for the return that can be expected from investing in a security without taking any risk. Even though it is a theoretically approved concept, it is hard to define in practice what the risk-free rate is and how it can be obtained. Both academics and industry have used government treasury bills as proxies for the risk-free rate, which are short-term loans that the government has taken for driving forward governmental projects.
**Alpha** $\alpha$ is the excess return that a security is yielding compared to the market in the FF3. $\alpha$ in general is used for evaluating the performance of an owned portfolio versus a given benchmark (usually the market) in order to measure if the portfolio is effective. If $\alpha$ is:

1. Negative, the security is not yielding the expected return, given the risk-profile of the portfolio.
2. Zero, the risk taken corresponds to the return of the security.
3. Positive, the security is performing better than expected given the risk taken.

Multiple managers are striving for a positive alpha, since it is an indicator of yielding high returns with a small amount of risk.

**Multiple Linear Regression** All the terms $\alpha, \phi_1, \phi_2, \phi_3, \epsilon$ are a product of multiple linear regression between the target return $E[R_a] - R_F$ and on the right side $(R_M - R_F), SMB, HML$. Each $\phi$ gives the slope for each of the factors and $\alpha$ is the sum of the intercepts from each linear fit between the factors.

**Factor Construction - HML and SMB** These factors represents the anomalies, which according to Fama and French CAPM fails to grasp. The SMB- and HML-factors are calculated by constructing six equally weighted portfolios, which happens as follows on the stock market at a given point in time $t$:

1. Select all stocks on the market and compute ME of those stocks
2. Sort all stocks according to their ME in descending order
3. Divide all the stocks into two portfolios by taking the median of the ordered stocks
4. The portfolio which has all the higher ME-stocks in it is the Big Portfolio (BP) and the portfolio with the smaller ME-stocks in it is the Small Portfolio (SP)
5. Sort the BP and the SP again separately by utilizing the Book-to-Market value of each stock in the portfolios in descending order
6. Divide both the BP and the SP each into three smaller portfolios at the 30% percentile and 70% percentile of the ordered stocks

The procedure is further visualized as follows in Figure 3 below:
Figure 3: How to create the SMB and HML factors in the FF3 model.

On the left side of Figure 3, the stocks are divided into SP and BP by their corresponding ME-values. Thereafter, the stocks within SP and BP are separately subdivided by into three smaller portfolios by their corresponding Book-to-Market value. After the stocks have been separated and categorized into different portfolios, the end result should now be six portfolios as follows:

<table>
<thead>
<tr>
<th>Table 2: Underlying portfolios in the SMB- and HML factors.</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% High BV/ME</td>
</tr>
<tr>
<td>40% Medium BV/ME</td>
</tr>
<tr>
<td>30% Low BV/ME</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

After the underlying portfolios have been constructed, it is possible to calculate returns for the portfolios, which are then utilized for calculating SMB- and HML factor values. The return of the underlying portfolios will be calculated based on equal weight, which means that each of the stocks in the portfolio are owned at the same ratio. This implies, that the return (for each of the portfolios) at a given point t in time will be as follows:

\[ R_P = \frac{1}{n} \sum_{a=1}^{n} R_a \]  \hspace{1cm} (10)

Where R is the return, n is the amount of stocks in the portfolio, subscript a stands for a stock in the portfolio and P is an arbitrary portfolio containing n assets. Based
on the returns from the six underlying portfolios, the SMB and HML factors will be calculated as follows:

- \( SMB = \frac{R_{S/L} + R_{S/M} + R_{S/H}}{3} - \frac{R_{B/L} + R_{B/M} + R_{B/H}}{3} \)

- \( HML = \frac{R_{S/H} + R_{B/H}}{2} - \frac{R_{S/L} + R_{B/L}}{2} \)

The factor values are calculated at certain intervals after the six portfolios have been constructed, and the factors values will be reset the next time the six underlying portfolios are updated. As can be seen from the factor construction, SMB is constructed by taking the average returns from the SMB-factor (Small Minus Big) and HML is constructed by averaging the returns for HML (High Minus Low). The SMB-factor, therefore, encapsulate the risk-factors in returns, which are related to Market-Value-of-Equity (size-anomaly) for the stock. The HML factor in turn is encapsulating the risk-factors involved with the Book-to-Market Value (value-anomaly).[14]

**Conclusion**  
Fama and French [14] found that despite the fact that company size and value are not the descriptive key variables, higher average returns on small stocks and high book-to-market stocks reflect unknown key variables that are able to price the undiversifiable risk in returns, left by CAPM model. Furthermore, according to Fama and French’s paper from 1995 [16], the findings show that weak firms with prevailing low earnings tend to have high BE/ME and positive slope on HML, and negative slope in case of strong firms with persistently high earnings. HML appears to capture the variation of the risk factor related to earnings performance. Coupled with SMB, there are two main conclusions that can be drawn: 1) that stocks with low long-term returns tend to have positive SMB and HML slopes and 2) stocks with low long-term returns tend to have higher average returns. In contrast, the stocks with high long-term returns tend to have negative slopes on HML and low future returns.[45] Experiments carried out using the FF3 regression model explaining returns of individual stocks found that domestic markets have a higher success rate in encapsulating the time-series variation in returns and has generally lower pricing errors than world factor models.[19] Another independent experiment of international factors supports this claim that domestic FF3-factor models perform better than international ones.[23] The performance of country specific FF3-models in Europe has yielded fairly good results.[1]
3 Reconstructing the Underlying Portfolios using Graph-Based Semi-Supervised Learning

In the previous chapter, earlier theory stated how the FF3 model emerged as a result from critique against previous models and frameworks, and how to build the FF3 model from scratch. Even though it is widely used in industry today, the model is not perfect. Since Fama and French state that the descriptive key variables explaining stock returns exhibit a high correlation in returns, this chapter will focus on how to rearrange the stocks in the underlying portfolios, in such a way, that it places stocks with similar return profiles in the underlying portfolios to further improve the explanatory power of the standard FF3 model. This chapter includes an extension of a proposed machine learning framework, including an underlying network structure and semi-supervised learning, which enables applications to various factor models. The focus will thereafter be on the rearrangement of stocks tailored specifically for the FF3 factor model, to unravel the descriptive key variables from the Fama French paper[14].

Machine learning can be summarized as programming computers, to optimize a performance criterion using example data or past experience. Machine learning applies to cases, where there does not exist an accurate model that can directly solve a given problem. Therefore, the application would be to give existing information about the problem to the computer and let the computer solve the problem.[2] For solving the problem, what is needed is features (D), that contribute to a certain outcome (Y). An example would be to classify if a person should be accepted for credit loans (Y), based on various features (D) about the person. In extreme cases, such as a person never paying debts, it would be an easy task for a human to rationalize the denial of the credit loan, due to knowledge about a history of not repaying debt. However, finding the fine line between granting and denying the loan would be a more rigorous and complex process for a human being, since there might not be 1) any earlier data of the person raising credit or 2) guarantees that the credit will be paid back. Therefore, successful machine learning approaches for credit scoring have been applied to rank credit customers.[9] In the example, a solution to the problem would have been to classify the customers as either grant loan or deny loan (2 classes). Depending on earlier knowledge about the output of the problem, there are three main categories of machine learning problems: 1) supervised learning, 2) unsupervised learning and 3) semi-supervised learning. Supervised learning is aiming to find a function, that can describe the relationship between the features (D) and the output (Y). In supervised learning, the output (Y) and features (D) are known, and the goal is to handle new incoming features (D\text{new}) that maps them to the respective output (Y\text{new}). Unsupervised learning on the other hand is trying to find structure in the features (D), without having any information of the output (Y) to assist. Since there are no output examples in unsupervised learning, it is also hard to evaluate the correctness of the model. Semi-supervised learning can be considered as a midpoint between supervised and unsupervised learning, where there are little known outputs (Y) but a lot of unknown outputs. Semi-supervised
learning makes use of both known outputs and unknown outputs. There are two ways of representing the output of the problem, which is either 1) classification or 2) regression. A classification problem is when the output belongs to a certain class, in the credit loan example, one class would be 1) grant credit, and the second class would be 2) deny credit. If the output would be continuous, then the output could for instance represent a probability function.

3.1 Graph Based Semi-Supervised Learning

Due to the expansion of content on the Internet and the increasing amount of unlabeled data, several semi-supervised learning (SSL) algorithms have been developed. By utilizing both labeled and unlabeled data, SSL tackles the problem of labeling unlabeled masses of data. On another line of research, graphs provide a natural way of representing data in multiple different domains. By combining graphs and SSL, graph-based SSL has reduced the amount of human effort and outperformed multiple state-of-the-art solutions in areas such as speech recognition, computer vision, natural language processing and multiple other areas of artificial intelligence. SSL is the midpoint between supervised and unsupervised learning, where a small fraction of the training set D is labeled and the rest of the training set is unlabeled. The goal of a SSL algorithm is to learn a function that is able to map the unlabeled data points to the labeled data points. Since the unlabeled data can not be mapped straightly from the labeled data points, SSL algorithms utilize preset assumptions to perform the mapping:

1. Smoothness assumption = If two points in a high-density region are close, then the output is also similar.
2. Cluster assumption = If two points are part of the same cluster, they are most likely to belong to the same class.
3. Manifold assumption = High-dimensional data can be expressed in a lower dimension

Different SSL algorithms assume one or a combination of the assumptions to hold. Graph-based SSL algorithms are a subset of SSL, which has received attention in the past. Furthermore, graph-based SSL algorithms are very attractive since: 1) multiple datasets can be expressed as graphs, 2) many of the graph-based SSL problems involves solving a convex objective function and 3) multiple graph-based SSL are scalable. Graph-based SSL utilizes (at least) the manifold assumption, where both the labeled and unlabeled data is within a low-dimensional manifold to be expressed as a graph. In the graph, the nodes correspond to a data-sample and the edges are weighted by the similarity to the neighboring nodes. There are two different types of graph-based semi-supervised learning strategies, which are: 1) inductive learning and 2) transductive learning. Inductive learning means that the model is being trained to label unseen nodes that may appear to the graph in the future. Transductive learning, which will be the focus of this thesis, means that the labeled
and unlabeled nodes are already given, and the problem is to label the unlabeled nodes without new ones appearing. As proposed by Subramanya and Pratim[53], the following steps needs to be included for a graph-based SSL approach: 1) graph construction (if no graph is given), 2) assigning labels to nodes from the initial sample and 3) infer labels on unlabeled nodes on the graph. What I propose, is an extended version of the framework presented by Subramanya and Pratim[53]. It takes into account the selection of initial labels, to incorporate graph-based semisupervised learning in the construction of various factor models. In the reconstruction of the FF3 factors, this translates to what can be seen in Figure 4 below:

![Figure 4: How to reconstruct the FF3 factors with Graph Based Semi-Supervised Learning.](image)

As can be seen in Figure 4, the proposed framework involves four steps as in the proposition mentioned by Subramanya and Pratim[53]. However, the steps in this framework will be to 1) first separate out the appropriate labels, 2) build the network and feed the labels to the network, 3) use graph signal recovery to recover signal values from the graph and 4) recalculate the factor models based on the resulting outcome from the graph signal algorithms. In this case, the recalculated factors would be the SMB- and HML factors as defined by the Fama and French paper[14]. The rest of the chapter discusses each step in detail.

### 3.2 Label Selection

In traditional settings of supervised classification, all outcomes (labels) are known to train a classifier. However, in real world data, labeled instances are usually hard to come by, and very expensive to acquire with respect to time and effort. Unlabeled data on the other hand is easily accessed in large volumes and the data is easy to store. Semi-supervised learning addresses the problem of scarcity in labeled data, by
combining: 1) a smaller amount of labeled data and 2) a larger amount of unlabeled data, to build better classifiers. The benefits of combining both labeled and unlabeled data are: 1) less human effort and 2) a higher classification accuracy, rather than using the traditional supervised classification framework. \[60\]

Graph-based semi-supervised learning is a subarea of semi-supervised learning, where the starting point is a graph with some nodes labeled and other nodes unlabeled. The edges between nodes are weighted and reflect similarities (or distance) between nodes. Underlying many supervised learning algorithms is the smoothness assumption, which states that nodes with large similarities (or small distances) belong to the same class. In semi-supervised learning, the smoothness assumption yields the algorithms to: 1) label effectively the unlabeled nodes with the labeled nodes as sources and 2) gives decision boundaries in low-density regions. \[26\][62] The smoothness assumption will be of high value, to extract the descriptive key variables, since it will ensure similar stocks being placed in the same underlying portfolio. The portfolios of the factors in the FF3 model are already known to belong to a portfolio (labeled) after the factor construction. For label selection in semi-supervised learning, it would be possible to take initial samples from the stocks and label them, to indicate which portfolio the labeled stocks belong to. To be coherent with the original FF3 model, samples would be taken from the stocks that most likely belong to an underlying portfolio. Below is a visualization of the sampling from the FF3 factor construction:

![Figure 5: Initial samples from the FF3 factor construction in Figure 3.](image)

As can be seen in Figure 5, each respective color indicates which underlying portfolio the sampled stocks will belong to as known labels. The positioning of the coloured boxes also highlights were the sampled stocks will be taken from, which is at the top or bottom percentiles depending on the portfolio. This will ensure that the stocks will with a higher probability belong to the respective underlying portfolio, rather than random sampling.
3.3 Networks

During the second step of Figure 4, a network needs to be constructed to connect the stocks with similar return profiles. This subchapter discusses what a network is, how it can be used and how it can be applied to the GB SSL solution.

Network Theory

A network is a simplified representation of a system, covering 1) the individual parts of the system and 2) connections between each individual part. There are many systems that are composed of individual parts, and are, therefore, representable as network structures. These systems are of interest, since it gives insight into the structure by investigating: 1) how the individual part is built up, 2) how the links are communicating with individual parts and 3) the pattern of connection between components. For instance, if there is a certain company on the market, it would be interesting to see interaction between key data of the company such as: stock price, size of the company, accounting data on a yearly basis, to understand the structure of the company. Another angle would be to see if the company has any competition or has logistics or services with other companies, to see how the company interacts with other companies on the market. Finally (as is done in the thesis), it would also be of interest to investigate patterns in the interaction between companies, such as stock price movements in the same or completely opposite direction than another company and to further analyze causes of similar or opposite movements.

Multiple academic domains such as 1) computer science, 2) biology and 3) social sciences, are using networks, and networks are widely used in today’s world, such as: 1) the Internet, 2) telephone networks, 3) power grid networks, 4) transportation networks and 5) delivery and distribution networks. The analysis of these networks will give insight into the structure of the systems, such as how critical an individual part of the network is or how to reach, as efficiently as possible, from an arbitrary point A in a network to an arbitrary point B. In case there is a behavioral pattern in the network, it is possible to make mathematical predictions of processes within a network. A network consists of two basic parts: 1) nodes and 2) edges. A node is a representation of an individual part in a system, such as a company in a financial network or a city in a geographical network. An edge describes the relationship between two separate nodes, such as a company buying its materials to make its product from another company, a data line between two computers or a road between city A and city B. In some cases, the edges have certain weights assigned to them called edge weights. Edge weights describe the connection strength between a connected node A and node B. For instance, when water flows through pipes, the water flow in the pipe is restricted by the area size inside the pipe. The pipe is, hence, only able to tolerate a limited amount of water that can flow through that pipe per second. The pipe can also be designed in two ways, were the water can either flow in one particular direction or in both directions. Hence, there exists two different types of graphs: 1) directed graphs and 2) undirected graphs. A directed graph is a graph where the edges have a defined direction like the water only being able to flow from a node A to node B, but not from node B to node A. Undirected
graphs on the other hand are graphs without a specified direction, such as a road network with two-way streets.

Analysis of networks is usually performed using computers, which means that the storage of networks needs specific data structures. The two most common ways of storing networks are in 1) list structure or 2) matrix structure. A matrix structure storing a network is called an adjacency matrix and a list structure storing a network is called an adjacency list. It is more common to store a network in an adjacency matrix when the network is small, since it is more convenient to perform fast analysis on an adjacency matrix. An adjacency list is respectively used when the network is large, due to the large amount of memory required to store a large network in an adjacency matrix. Depending on the size and the use case of the network, one structure is preferred above the other.

In this thesis, we will focus on the adjacency matrix. The adjacency matrix represents the nodes and its connections within the network as an nxn matrix, where the index numbers of the matrix represents the nodes and the values in the matrix the connections. The logic behind storing the network within an adjacency matrix is to assign each node a unique index number in the matrix from 1 to the number of nodes noted as n. Since each index corresponds to a unique node, it is possible to represent each edge in the matrix by setting a nonzero number where there exists an edge between two nodes. For instance, in an undirected graph, if node number 1 and node number 3 have an edge between them, the matrix Network on index Network(1,3) = 1 and Network(3,1) = 1. In case there is no edge between two nodes, one strategy is to set those adjacency matrix-indexes to zero.[41]

The incidence matrix of size n x m is another representation of the network, where the columns are the nodes from 1 to n, and the rows are the edges from 1 to m. The logic behind the incidence matrix is to assign each of the nodes’ index number on the columns, and list the edge connection on each row. One row will represent one edge, where the indication of an edge will be set to 1 (or corresponding edge weight) under the column(s) where the edge connects the nodes and the other column values will be set to zero. Notable in the incidence matrix is that the order of listing the edge-connections is irrelevant. Since the thesis will only cover the use case of an undirected network stored in matrix form, that case will be covered in more detail. The thesis is restricted to not cover special cases of graphs, such as graphs containing self-edges (nodes which have edges to itself) or cycles (closed loop of edges).[41] How the previous stated theory can be utilized for measuring similarities between stocks, will be discussed more in detail in the next subchapter.

**Stock Correlation Network**

As stated in the previous subchapter: 1) networks are used for representing a system, 2) different types of networks are used for different use cases and 3) ways of storing a network in a data structure. The following chapter will cover how to model the market as a network using stocks and their respective returns, i.e. a stock correlation network. A stock correlation network consists of nodes that represent stocks on the market, and edges that represent the correlation value between the stocks. Building
a stock correlation network is done by following these four steps: 1) select the desired time series data, 2) calculate the correlation value between each pair of stocks from step 1, 3) build the correlation matrix using the results from step 2, 4) transform the correlation values into a distance, to be able to apply network algorithms on the network.[34] As stated earlier, networks consist of nodes and edges, where weights can be on/off or have weights assigned to them. In economics, one way of looking at how similar (or dissimilar) returns in stocks are, in a certain period, is to utilize the Pearson correlation coefficient (PCC) defined below:

\[
\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sigma_x \sigma_y} \tag{11}
\]

Subscript x refers to the return data for one stock, subscript y refers to the return data for another stock, the bar indicates the mean value and \(\sigma\) the variance. PCC will have values ranging between \([-1,1]\], and the values should be interpreted as follows:

\[
\rho_{xy} = \begin{cases} 
1, & \text{stock returns are moving in the same direction} \\
0, & \text{stock returns are not moving in a certain pattern} \\
-1, & \text{stock returns are moving in opposite direction}
\end{cases}
\]

As can be seen in the conditional statement above, the closer the PCC is to one of the (extreme) values, the more information can be extracted about the movements between two arbitrary stocks. By defining the stock market as a system with individual stocks as nodes, it is possible to compute the correlation coefficient on the return data between all individual stocks on the market, and hence construct the correlation matrix defined below:

\[
W_{i,j} = \begin{pmatrix} 
\rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,j} \\
\rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,j} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{i,1} & \rho_{i,2} & \cdots & \rho_{i,j}
\end{pmatrix} \tag{12}
\]

Some properties with respect to the correlation matrix are that: 1) it is a square matrix, 2) it is a symmetric matrix 3) the values will be from \([-1,1]\]. Since the matrix is both a square matrix and symmetric, it can be treated as an undirected network. However, to treat the matrix as a network, the edge weights need to be converted to a distance to measure how close or far away each node is. To measure how close or far away each node is, the distance between nodes needs to be a metric. A metric is defined by holding the following four properties: 1) all distances are nonnegative, 2) the distance of an object to itself is zero and distinct objects are never at distance zero, 3) the distance between two objects is the same in both directions and 4) the distance satisfies the property that detours are longer (triangle inequality).[11] Since PCC violates the properties of a metric, a transformation is necessary.[34] To convert PCC to a distance, the metric properties have to be satisfied. One such distance is the absolute correlation distance defined below:

\[
\sin(\theta) = \sqrt{2(1 - \cos \theta)} = \sqrt{2(1 - \rho_{xy})} \tag{13}
\]
Where a value closer to zero means that the stocks are moving in the same direction. On the contrary, a \( \sin(\theta) \)-value closer to the value two, means that the stock returns are moving in an opposite direction. The purpose of transforming the data into a metric is to be able to apply different processing and analysis techniques on the network in a later stage. For instance, Dijkstra’s algorithm for finding the shortest path from node A to node B in the networks (geodesic distance) or creating a minimum spanning tree (MST). After the stock correlation network has been built, the outcome is a fully connected graph with varying edge strengths between nodes. Due to the fact that historical stock returns contain a lot of noise, as has been discussed as one of the issues with finance and shown empirically by Tse et al.[55], only the strongest links should be kept in the stock correlation network. Furthermore, Mantegna[34] states that such graph topology gives a solid description of how the financial markets are built, and which economic factors influence certain groups of stocks. Such graph structure is obtained by utilizing information present in time series of stocks. Empirical research conducted by Mantegna shows that stock prices carries valuable and detectable economic information, which presents a window into understanding how certain types of stocks influence present market dynamics. On the other hand, for the propagation algorithms to work properly, one criterion is that the network should be connected (in other words, there are no unreachable nodes in the system). To ensure that the network is connected, one possibility is to find the minimum spanning tree (MST)[47][31] from the network. The MST is a subset of a connected and undirected graph, which has the following properties: 1) the MST is connected, 2) the MST is undirected, 3) the MST has no cycles and 4) the MST has a minimum total edge weight. These properties ensure that the propagation algorithms will assign labels to every stock in the network, and will keep the strongest links between each node intact. Since computing the exact MST is NP-complete, it is not feasible to calculate the exact MST for a stock correlation network on a large market. To solve this issue, one approach is applying a greedy algorithm on the MST, where two popular algorithms exist: 1) Kruskal’s algorithm[31] and 2) Prim’s algorithm [47]. Since both Kruskal’s and Prim’s algorithm can be solved in polynomial time, both methods can be used. The output from Prim’s algorithm will produce a tree, where each of the nodes will have one outgoing edge to another node. An example of a resulting tree from Prim’s algorithm can be seen below in Figure 6:
Figure 6: Minimum Spanning Tree on the US stock market as a stock correlation network 30.06.2000.

Even though the algorithm has ensured that each pair of node is connected, it does not preserve other strong links inside the network. Problems are that: 1) the network loses vital information about the network structure, and 2) the propagation algorithms need additional connections to work properly. Otherwise, the result from the algorithms will be exclusively dependent on the positioning of initial labels and labels are not propagated efficiently. To adjust the amount of connections in the graph and retain as much information as possible, one approach is to use the winner-take-all approach on the network. The purpose of winner-take-all is to spare all the connections above a certain threshold value, and to shut down the other connections.[55] There are several ways to threshold the network, but the primary threshold strategies in this thesis are: 1) network density and 2) largest node degree.

The network density is the number of existing edges divided by the maximum amount of possible edges in a network, which is defined as:

$$\rho_{\text{NeDe}}(G) = \frac{2M}{N(N-1)}$$  \hspace{1cm} (14)

Where $M$ is the amount of existing edges in a network, $N$ is the number of nodes and $\rho_{\text{NeDe}}(G)$ is the network density, describing how connected the network is. If no edges would exist in the network, the network density would be zero. If every node has edges to every other existing node in the graph, the network density would be 1. The node degree $d(u)$ is the number of (outgoing) edges from node $u$. The largest node degree is defined as follows:

$$\rho_{\text{NoDe}}(G) = \max_{u \in G} \left( \frac{d(u)}{N-1} \right)$$  \hspace{1cm} (15)
Where $\rho_{\text{NoDe}}(G)$ is the ratio of connections, belonging to the node with the largest degree. The difference between the two threshold measurements is that $\rho_{\text{NeDe}}(G)$ focuses on the overall degree between nodes in the network, whereas $\rho_{\text{NoDe}}(G)$ focuses only on the node with the largest degree. Below is a visualization of how the winner-takes-all set its threshold value on a stock correlation network.

![Graph Threshold](image)

Figure 7: Measuring threshold values to a stock correlation network using $\rho_{\text{NoDe}}(G)$ and $\rho_{\text{NeDe}}(G)$, with initial requirements keeping the edge density at maximum at 0.5%, and the most connected node at a maximum degree ratio of 2%.

As can be seen in Figure 7, the edge density increases as we allow weaker links to be connected in the graph. Also, the degree for the most connected node increases at almost the same pace. By keeping the edges with a higher weight $(\max(\rho_{\text{NeDe}}, \rho_{\text{NoDe}}))$, it is possible to merge the subset with higher weights than the threshold with the previously computed MST as follows:

$$G(\text{final}) = G(\text{MST}) \cup G(\text{Threshold})$$  \hspace{1cm} (16)

Where $G(\text{MST})$ and $G(\text{Threshold})$ are subsets of the original network. In Figure 7, the corresponding threshold would be the threshold for $\rho_{\text{NoDe}}$, as its threshold value enforces a higher correlation ($\rho$-value) between stocks. After the $G(\text{MST})$ and $G(\text{Threshold})$ subsets have been computed, both subsets are overlapped and the resulting network can then be injected with stocks selected from the underlying portfolios. Below in Figure 8 is a visualization of the resulting graph with labels injected.
Figure 8: A stock correlation network, merged as a combination of MST and winner-take-all, injected with labels.

As can be seen in Figure 8, the black nodes are stocks that do not belong to an underlying portfolio, whereas the coloured nodes represent stocks that have initially been labelled to one of the six underlying portfolios. The next step in reconstructing the six portfolios, in the Fama French model, is identifying the labels of unlabelled nodes, using graph signal recovery methods.

3.4 Graph Signal Recovery

As stated earlier, the initial graph has been constructed and injected with labels. The next step is to label all the unlabeled nodes in the stock correlation network to (re)construct the portfolios in the FF3 model. Reconstructing the portfolios is performed by utilizing graph signal recovery. The goal of graph signal recovery is to solve labels or signal values on the unlabeled nodes $Y_u$ by utilizing 1) the signal values of the labeled nodes $Y_l$ and 2) the constructed graph $G$. The algorithm will then utilize the preset assumptions of GB SSL discussed earlier, to propagate the signal values. Finally, the output of the signal recovery algorithm will be an estimate of the signal values $\hat{Y}$. Signal recovery for graphs can be expressed as a convex optimization problem, utilizing a trade-off between 1) the bias of the graph signal and 2) the smoothness of the graph signal defined below[35]:

$$\hat{Y} = \arg\min( E(\hat{Y}[:]) + \lambda R(\hat{Y}[:]) )$$  \hspace{1cm} (17)

$E(\hat{Y}[:])$ Measures how accurate the model is, whereas $R(\hat{Y}[:])$ measures how smooth the predicted signals are. A high value of $E(\hat{Y}[:])$ indicates that the model has a lot of errors, whereas high values of $R(\hat{Y}[:])$ indicates that there are large jumps between the graph signals. The parameter $\lambda$ tunes the weight between the bias $E(\hat{Y}[:])$ and
the variance $R(\hat{Y}[:])$ of the recovered signal values. In this thesis, two graph signal recovery or propagation algorithms will be covered: 1) label propagation and 2) sparse label propagation.

**Label Propagation**

Label Propagation is one of the earliest graph signal recovery algorithms, which has had several successes in multiple domains[62][35][30]. The algorithm aims at minimizing the variance between the signals, and hence at keeping the graph signal smooth.[62] The algorithm takes as input a connected and weighted graph $G$ and a subset of labelled nodes $Y_l$, and produces as output an estimate of the graph signal $\hat{Y}$. Smoothness of the graph signal is achieved by averaging the signal value of nearby nodes, while keeping the initial labelled nodes $Y_l$ unchanged. By utilizing the convex optimization framework discussed in the previous subchapter, the algorithm sets the bias term $E(\hat{Y}[:]) = 0$ and aims to minimize the variance $R(\hat{Y}[:])$. The variance is measured by utilizing the $l_2$-norm between the nearest neighbors’ signal values $\hat{y}[i]$ of the node, and multiplied by their corresponding edge-weights $w_i$. The minimization problem is defined below:

$$R(\hat{Y}[:]) = \sum_{(i,j) \in e} W_{i,j} ||\hat{y}[i] - \hat{y}[j]||^2_2$$

and can also be expressed in the following form by utilizing the graph Laplacian $L$ as:

$$R(\hat{Y}[:]) = \hat{Y}^T[\cdot]L\hat{Y}[\cdot] = \hat{Y}^T[\cdot](D - W)\hat{Y}[\cdot]$$

By minimizing with respect to $\hat{Y}[:]$, the closed form solution would be

$$\hat{Y}^T[\cdot] = D^{-1}W$$

Hence, the LP algorithm can be written as:

Algorithm 1 Label Propagation

1: function LABEL_PROPAGATION(W, Y, e) \triangleright Where W - Weighted Matrix, Y - Initial labels, e - Stopping criterion
2: $D_{ii} = \sum_j W_{ij}$
3: $\hat{Y}^0 = Y$
4: \hspace{2em} $k = 1$
5: \hspace{2em} \textbf{while} $\nabla \hat{Y} > e$ \hspace{2em} \textbf{do}
6: \hspace{4em} $\hat{Y}^{k+1} = D^{-1}\hat{Y}^k$
7: \hspace{4em} $\hat{Y}_l^{k+1} = Y_l$
8: \hspace{4em} \hspace{2em} $k = k + 1$
9: \hspace{2em} \hspace{2em} \textbf{end while}
10: \hspace{2em} \textbf{end function}

Label propagation has had several successes in various areas. However, keeping the initial signals unchanged might influence algorithm performance (if the initial...
signal selection is biased). Other issues that might affect the performance of label propagation is low tolerance against drastic changes in signal values of nearby nodes.

**Sparse Label Propagation**

One recently developed propagation algorithm is sparse label propagation, proposed by Jung et al.[26]. The signal values are modeled using the total variation of graph signals, which has shown to model graphs containing abrupt changes better than the Laplacian quadratic form[8] utilized in label propagation. The optimization problem can hence be written as:

$$R(\hat{Y}_{TV}) = \sum_{\{i,j\} \in \epsilon} W_{i,j} ||\hat{y}[i] - \hat{y}[j]||_1$$

As can be seen in Equation 21 above, the optimization amounts to solve the $l_1$-norm whereas label propagation aims to minimize the $l_2$-norm. The $l_1$-norm in sparse label propagation results in better results for graphs with abrupt changes in graph signals than the label propagation algorithm. The difference can be seen below from experiments conducted by Jung et al.[26]

![Figure 9: Label propagation and sparse label propagation models on a data graph in [26].](image)

As can be seen in Figure 9, sparse label propagation yields a better estimate of the signal values than label propagation. When to use one algorithm instead of the other algorithm, depends on the underlying network structure as both have their own core strengths and weaknesses. Label propagation will perform better on graphs with a smooth graph signal distribution, whereas sparse label propagation will perform better if there are clear jumps of signal values in the graph. By applying both graph
signal recovery algorithms on the injected network in Figure 8, the following result was obtained in Figure 10.

![Graph Signal Recovery Algorithms](image)

**Figure 10**: Stock labels recovered by the graph signal recovery algorithms label propagation and sparse label propagation, as a result from the stock correlation network in Figure 8.

With the results from Figure 10, it is now possible to divide the stock into the underlying portfolios according to the FF3 model. Since each stock is now labelled to belong to one of the six portfolios in the FF3 factor model, it is now possible to proceed to Equation 10 and use newly labelled stocks to reconstruct the six underlying portfolios with labels from 1) label propagation and 2) sparse label propagation.

This part of the thesis presented an innovative way of reconstructing the underlying portfolios for the original, economic theory based factor model, enabling a reconstruction of various factor models using different graph-based semi-supervised learning algorithms. Thereafter, the chapter has revised the particular structure of the underlying graph, and presented the algorithms used in the empirical research part of this thesis. As a result, the underlying portfolios can now be reconstructed based on stock movements, in an attempt to capture the descriptive key variables mentioned in the Fama and French article[14]. The study then proceeds with an attempt to increase the explanatory power of the model in relation to stock returns, in comparison to the explanatory power of the standard FF3 model, where the factors are constructed in the manner initially proposed by Fama and French.
4 Research Methods

The research method used in this thesis is a combination of empirical and quantitative research, utilizing a database and mathematical analysis-software as main tools in conducting the experiments. The tools utilized were MatLab for data analytics and a MySQL database for storing intermediate or final results. The data for the experiment was imported from the Wharton Research Data Services (WRDS), which is utilized for quantitative research and hence has become the standard for conducting research within business intelligence.[58] Finally, the regression results were compared using R2-adjusted as a quantitative measure for explanatory power.

From the WRDS, the following databases were used: CRSP, CompuSTAT and treasury bill data from Ibbotson Associates. More specifically, from these databases the following information was imported: 1) Annual balance sheet data (funda), 2) monthly market price data (msf) and 3) risk-free rate from Ibbotson Associates. Since the tables were all in .sas-format, a Python script was implemented for storing them into the MySQL database. When building the Fama French three factor models in this thesis, the model construction was identical to the Fama-French paper from 1992 according to the following guidelines: 1) all financial firms were excluded from the analysis, 2) stocks from the following exchanges were accepted: AMEX, NASDAQ and NYSE and 3) price, shares outstanding and book to market value had to be defined for the stock. When calculating the size for each stock, the Market-to-Equity value in June year t was used. However, when the Market-to-Equity value incorporated in the Book-to-Market ratio was calculated, the market value for December year (t-1) was used. After each ratio was computed and when the stocks were incorporated in the underlying portfolios, the model was reset and recalculated once a year in June (see Figure 3). After the underlying portfolios were defined, values for SMB- and HML factors were calculated once a month based on the equal weighted return from each portfolio. The return for the underlying model was calculated until the underlying portfolios were reconstructed again, in June year (t+1). The specified timeframe for the experiment was between 2000-2015, which means that the underlying portfolios were reconstructed 15 times in total.

Since there are a lot of steps and components in the FF3 model, the process is easier to comprehend by dividing it up into three separate events: 1) constructing the six underlying portfolios each year, 2) calculating the value of the factors using the six portfolios and 3) computing the model value based on the factors. Since the network models are also incorporated into the construction of the FF3 model, the process is divided into three separate Figures. In Figure 11 below can be seen how the underlying six portfolios and the values of the factors were computed.
In Figure 11, the blue boxes represent MySQL tables where results are stored and/or utilized for further computations. The transparent boxes represent heavier computations performed in Matlab and the black boxes represent the relevant output from the process in Figure 11. Since there is no defined table in the Figure were the relevant output was stored due to limited space, it will be referenced to as “Standard Model Components” hereafter. Starting from the left side of Figure 11, the data is located in three separate tables. The “Treasury Data”-table contains the value for the 1 month Treasury-Bill, which was used as a proxy for the risk-free rate. The balance sheet data contains yearly data from corporations’ balance sheets, which was utilized for calculating the Book-Value for each stock. The monthly return data covered the rest of the data needed for the FF3 model. To simplify the management of data, the three tables were merged into the table 'Table for Factor Construction' utilizing a MySQL-procedure. Thereafter, from 'Table for Factor Construction', both the risk-free rate and the market value were straightforward to compute and store into 'Standard Model Components'.

When calculating the value for the SMB- and HML factors, Matlab had to be used to handle the process for constructing the six underlying portfolios. Matlab handled the construction of the underlying portfolios and calculated factor values each month, sending the values to the database, which stored them into 'Table for Factor Construction'. Since the network models require sampled stocks and their corresponding underlying portfolio, each year samples were taken from the portfolios to be used later by the graph signal recovery algorithms. The number of samples were either 5%, 10%, 15% or 20% of the total size of the underlying portfolio, where the selection process was targeted against stocks belonging most certainly to the respective underlying portfolio (as illustrated in Figure 5). Since the sample size may affect the outcome of the graph signal algorithms, its effect on the result
is investigated further in this thesis. When constructing the network models, the approach is similar architecture-wise on a higher level, but involves more steps than the standard model. Below is a visualization of how the underlying portfolios and the factor values were calculated when using the network models:

![Architecture for constructing the factors in the network models.](image)

**Figure 12:** Architecture for constructing the factors in the network models.

As can be seen in Figure 12, the same table 'Table for Factor Construction' as in the standard model was utilized. The first step is to import stock returns from the 'Table for Factor Construction', to construct the stock correlation network. After the stock correlation network had been constructed, samples extracted from the standard model were injected into the network. From here, the graph signal recovery algorithms resolved labels on each unlabelled stock using both the stock correlation network and the initial labels. Thereafter, the six underlying portfolios were constructed using the obtained labels from the graph signal recovery algorithms. From this point forward, everything is coherent with the standard model. However, since the timeframe in calculating the correlation between stocks can vary and, therefore, have an impact on the result, separate correlation lengths were incorporated to investigate long-term versus short-term effects into the construction of the stock correlation network. After all the components in the factor construction for both the standard model and the network model have been stored into the database, the final step is building the resulting FF3 regression model and comparing the model to stock returns. Below is illustrated how the FF3 models were built and compared against stocks on the US stock market.
Figure 13: Architecture for constructing the factors in the standard FF3 model.

Figure 13 shows how data is transferred to the table “Factor Construction Results” containing the SMB factors, the HML factors, market return and risk-free rate required for the FF3 model as a result from the procedures in Figure 11 and Figure 12. Each term in the FF3 model is thereafter imported to Matlab, where the model will be built and fitted to the return of stocks on the market. In order to measure the performance from each model, diagnostics from each resulting model were resolved, such as: R2-values, mean square error, Ljung-Box test, Engle-Granger test and Jarque-Bera test. The diagnostic values investigate the following information regarding the regression model:

- **R2** is a goodness-of-fit test, which investigates how large portion of the variation can be explained with the model. The value ranges between [0,1], and the closer the value is to 1, the better it can explain the variance of the model.

- **R2 Adjusted** is measuring the same as R2, but includes a penalty to the number of variables in the model to prevent overfitting. Therefore, due to the penalty, R2 adjusted will always have a lower value than R2.

- **Mean Square Error** is yet another goodness-of-fit test, which investigates how much the model in average deviates from the correct values. Unlike R2 and R2 adjusted, MSE ranges (in theory) between [0,inf] and the closer the value is to 0, the better.

- **Ljung-Box** measures randomness in a time-series, where the null-hypothesis means that the data is independently distributed. If the null hypothesis gets rejected, there exists serial correlation in the data. The Ljung-Box test is used to investigate if the error between the model and the return of the stocks is random. Therefore, the result could indicate white noise.
• Engle-Granger is a cointegration test, which investigates if the model parameters are not i.i.d. If they are not i.i.d., parameters can be expressed as a linear combination of other parameters. In the experiment, the point of interest is to ensure that the factors are not cointegrated.

• Jarque-Bera is a goodness-of-fit test, which investigates if a time-series has an underlying probability distribution matching a normal distribution. The JB-test will be used as a measurement between the error of the model and return of the stocks, to investigate if the result indicates white noise.

We have discussed the underlying process behind obtained results in the chapter that follows. To summarize, we’ll go through parameters that were used in the experiment and justify the choice of them. During the network construction, the network was restricted by keeping the node degree below 0.5% and the size of the largest hub below 2% of the total amount of (possible) connections. This is done to ensure that the strongest links remains intact without being susceptible to noise. The degree of the largest node was set to remain at a low level, to avoid highly connected nodes to distort the results. Within the graph signal processing algorithms, sparse label propagation was set to a $\lambda$-value of 1. The $\lambda$ value was chosen based on results between label propagation and sparse label propagation, experimental observation in the construction of results, as well as experimental results from Jung et al.[26] and Mara[35], where $\lambda = 1$ produced different outcomes compared to the label propagation algorithm. R2 Adjusted is used as a measurement of explanatory power in the regression models, since it is an academic standard within conducting econometric research[kalla]. MSE is used when investigating interclass (standard, label propagation and sparse label propagation) differences similarities, since there were no extreme values distorting results from the MSE.
5 Results

From here onwards, we will discuss the obtained results by applying the research methods described in the previous chapter. Since 1) the stock correlation network builds 2) the underlying portfolios, the underlying portfolios create the factors, and the factors are part of the FF3 model, the results will be investigated in the preceding order. The first subsection compares how the threshold was set for the stock correlation network at different correlation lengths and how it is reflected on market behaviour. Thereafter, the underlying distribution of stocks from the graph signal algorithms will be compared on different sample sizes. Finally, similarity of underlying stocks in the resulting portfolios from the graph signal algorithms, will be compared against resulting portfolios from the standard FF3 model. The third subsection investigates how similar the SMB- and HML-factor values produced by: 1) the standard model, 2) label propagation and 3) sparse label propagation are, and discusses differences between the results. The third subsection compares the regression results produced by all three models, the level of explanatory power each model produces and whether statistical methods can enhance the level of explanatory power.

5.1 Stock Correlation Network and the Six Underlying Portfolios

In the first phase of creating the stock correlation networks using graph-based semisupervised learning, a stock correlation network was constructed. This network contains nodes that correspond to stocks, and edges corresponding to correlations between stock returns. The network is firstly constructed by using stocks on the US stock market between 2000-2015, whereafter the rest of the connections are shut down, except connections from the MST and the winner-take-all approach discussed earlier. The time period governing the amount of records taken into account is set by the parameter correlation length. Possible values for the parameter 1 year, 3 years, 5 years and 10 years. Stocks that do not hold a return history equal or longer to the correlation length are automatically pruned away from the network. Since each year 4 different networks were constructed, the number of networks constructed in total were 16 years * 4 correlation periods = 64 networks. Below is a visualization of how the threshold values changed for correlation lengths during different years.
Figure 14: Threshold values for each network in the FF3 factor construction.

Figure 14 presents how the threshold value changed for each of the networks (markers) using correlation length of 1 year, 3 years, 5 years and 10 years. Figure 14 shows that the shorter the length of the time series, the higher the threshold value for the graph. The increase in threshold value on shorter correlation lengths can be explained by a higher number of stocks in the network, since longer correlation lengths mean a higher number of stocks will be pruned from the network. Another observation would be the plunge around 2007-2008 for correlation lengths: 1 year, 3 years and 5 years, where the Lehman crisis could affect the result. Finally, the threshold value for 10 years increases steadily with time, which suggests stocks with a longer history are becoming more densely connected on the stock market.

Next, we will investigate the behaviour of the underlying portfolios. The construction of the underlying portfolios is performed by utilizing the resulting six portfolios from the standard model. From the portfolios, samples of stocks were taken and labelled to the portfolio they were taken from. Thereafter, the labelled stocks were injected into the graph, followed by the graph signals algorithm resolving the rest of the unlabelled nodes. In Figure 15 below, we will look at the relative size of the portfolio, to investigate the label-distribution from the graph signal recovery algorithms. First, we will investigate the relative size of the standard model, followed by the relative size of the graph signal algorithms at different sample sizes. The distribution is computed by counting the number of stocks in the underlying portfolio divided by the number of stocks in the market at time t. To avoid repetition, each color in the following figures refers to one of the six underlying portfolios in the standard FF3 model. On the x-axis is the time-interval from 2000-2015 with monthly observations, and on the y-axis is the relative size. Next to the labels on the right side of the plot, is the average size of the portfolio presented.
Figure 15 highlights the distribution of stocks in the standard model, being either 15% or 20% of the total market size. There are small bumps in the relative sizes, which is due to new stocks entering the market during that month. The sudden drop after the bump is due to the portfolio(s) being reset in every year June with new stocks.

Below is presented how the initial sample size affects the size of the portfolios. Since each sample size incorporates four different stock correlation networks, the relative size of the portfolio is presented as an average between the networks. Inside the portfolios, each portfolio is built on the presumption that stocks with similar returns are contained with stocks having a similar return profile. Due to this fact, the ratio will describe two separate cases: 1) a high ratio means that there are a large amount of stocks with a similar return profile and 2) a low ratio means that the return profile of these stocks were a rarity in the network. Each plot will contain the same information mentioned above, but with the addition of two subgraphs. The top subgraph represents the relative size using label propagation, and the subgraph below represents the relative size using sparse label propagation.
As can be seen in Figure 16, there is a clear difference in both algorithms regarding the stock distribution compared to the standard model. With a sample size of 5% including both label propagation and sparse label propagation, the big neutral portfolio and big growth portfolio contain a larger amount of stocks with similar return profiles in the network. On the contrary, the big value portfolio and small growth portfolio contain a smaller amount of stocks, which means that the return profiles were a rarity in the network. Comparing both algorithms in the 5% case, sparse label propagation contains a higher variance in the number of underlying stocks in the portfolio. The most likely cause for this higher variance is that the algorithm converges to different outcomes each time the six underlying portfolios are reinitialized. Technically, the sample size might be low enough to alter the outcome of the model drastically, or, alternatively, there might be a sudden change in market dynamics during that period of time. When investigating underlying stocks in the small stock portfolios and big stock portfolios, the outcome is more stocks in the big stock portfolios than in the small stock portfolios. For the label propagation algorithm, the ratio is 65% big and 35% small, whereas for sparse label propagation the ratio is 71% - 29%.
Figure 17: Relative portfolio size of the underlying portfolios for the label propagation-and sparse label propagation algorithm for a sample size of 10%.

By comparing Figure 17 to Figure 16, we can observe some similarities and differences with a 10% sample size. One of the differences, for instance, is that the 10% sample size does not include as much variance between the reinitializations of the underlying portfolios. Another difference, is a more even distribution of stocks to the underlying portfolios. One possible explanation for these differences is the increase in sample size, guaranteeing a higher number of stocks in the portfolios and more starting positions for the nodes to propagate their signal values. However, as can be seen in Figure 16, the small growth portfolio is not able to find similar return profiles with the 10 percent sample size. The number of stocks (after graph signal recovery) increases by 2%, over the number that was initially assigned to the portfolio. The big stock portfolios have a higher number of stocks with similar returns, than the small stock portfolios in Figure 17. For label propagation the ratio is 61% big stocks and 39% small stocks, and for sparse label propagation the result is 65% big stocks and 35% small stocks.
Figure 18: Relative portfolio size of the underlying portfolios for the label propagation- and sparse label propagation algorithm for a sample size of 15%.

By further increasing the sample size to 15 percent, the variance of the return contribution lowers and the distribution of stocks in the portfolios gets more evenly distributed. There are also small dissimilarities between the resulting ratio of the algorithms, which means that both algorithms are converging to find stocks for each underlying portfolio. The gap between big stocks and small stocks is getting smaller for both algorithms, though there still is approximately a ratio of big stock portfolios with 60% of the market and small stock portfolios with 40% of the market. The small growth stock still has issues finding stocks with similar returns and remains the smallest portfolio.
Figure 19: Relative portfolio size of the underlying portfolios for the label propagation- and sparse label propagation algorithm for a sample size of 20%.

Finally, by labelling 60% of the stocks on the market, the result in Figure 19 is achieved. Comparing Figure 19 to Figure 16, Figure 17 and Figure 18, the sample size of 20 percent achieves the smallest variance between the portfolio reconstructions and a more evenly distributed amount of stocks in the portfolios. The big portfolios still have a higher number of stocks in their portfolios, where the label propagation algorithm amounts to 56% big stocks and 44% small stocks. For sparse label propagation, the corresponding ratios are 58% and 42%. The small growth portfolio still remains the smallest portfolio, either remaining at the initial sample size or increasing its number of stocks from 10% to 11% after graph signal recovery. Comparing the results obtained from graph signal recovery, sparse label propagation had more variance in its distribution of stocks than label propagation, when sample size were at a lower rate. By increasing the sample size, the ratio converged to similar results for both algorithms and the variance decreased. The result is logical, since increasing the initial number of labelled stocks makes it harder for certain labels to “dominate the network”. A recurring pattern in the return contribution of stocks was that big stocks had a higher number of stocks throughout the analysis. The results indicate that the return profile of big stocks (in line with the FF3 factor construction) concerns a larger group of stocks on the market. Another discovery is that the small number of stocks in the small growth portfolio, which throughout the analysis had smaller numbers of stocks inside the portfolio. The result indicates that the nodes with small stock labels got scattered around the graph on the sides of the graph, surrounded by small degree nodes. This could indicate that: 1) a small stock return profile is a minority or 2) the return profile is too complex for a correlation analysis to describe its relationship with other stocks. The ratio will have an impact of the outcome of the factors and their values, since each portfolio is considered equally. Due to this,
averaging the return on a smaller number of stocks will be subject to noise, due to the low number of stocks. Another issue is that it is known how the stocks have been distributed by the graph signal recovery algorithms, but not which stocks the portfolios contain. In Figure 20 below the coherence between the standard model and the network models (in average over correlation length and sample size) can be seen.

![Figure 20: Ratio of same stocks in the underlying portfolios for the network models.](image)

As can be seen in Figure 20, the x-axis describes the portfolio each category belongs to, with the corresponding color in the labels seen on the right side of the graphs. On the y-axis, is the ratio of similar stocks in the underlying portfolios from the standard FF3 model, compared to the network models. The top graph describes the similarity ratio between the standard FF3 model and label propagation algorithm, and the bottom graph describes the similarity ratio between the standard FF3 model and sparse label propagation. Both algorithms produce similar ratios of coherence with the standard model, but label propagation has more varying similarity with the standard model between portfolios, whereas sparse label propagation has more equal ratio of similarity throughout the underlying portfolios. The value stocks have larger similarities to the standard model than the growth stocks. The difference between value stocks and growth stocks gives an indication that there is a relationship between similar returns and their corresponding Book-to-Market value, where stocks with a high Book-to-Market value have similar return profiles. Growth stocks from the graph signal recovery models do not have a high similarity ratio with the standard model, which might give an indication of the network models describing another risk involved in growth stock that is not reflected when using Book-to-Market as a metric for encapsulating risk.
5.2 The SMB- and HML-factors

Previously, we have discussed the underlying portfolios and their differences in each of the model categories. Since the factors are built upon the return profile from the underlying portfolios, the portfolios will also affect the value of the SMB- and HML factors. The SMB- and HML corresponding factors produced by the three model categories will be the difference between the models. Next, we will compare factor values produced by the standard model, label propagation and standard label propagation by measuring the mean square error between the factors. Below is a comparison between the standard model and the network models.

![SMB Factor - Similarity between Standard Model and Label Propagation](image)

![SMB Factor - Similarity between Standard Model and Sparse Label Propagation](image)

Figure 21: Mean square error for the SMB-factor between the network models and the standard model.

Figure 21 describes how the values of the SMB factor differ from the standard model. On the x-axis are sample sizes and on the y-axis are the correlation lengths. On the z-axis is the mean square error between the SMB values from the standard model versus the graph signal recovery model. The top graph describes the similarity of the SMB-factor between the standard model and label propagation, and the bottom graph describes the similarity of the SMB-factor between the standard model and sparse label propagation. The color of the surface represents the value of mean square error between the standard model and the graph signal recovery model, with respective values for the colors on the right side. As can be seen in the figure, the value steadily decreases the higher the sample size from the underlying portfolios is. A higher sample size will automatically guarantee that at least the sampled percentage will be the same as in the standard model, which is why it influences the similarity in the factor values. Inspecting the correlation lengths, the similarity is higher at longer periods of correlation. Since the SMB factor is heavily influenced by the size of the company and its corresponding stock (and larger firms tends to be
stable and longer on the market), longer correlation lengths will prune away links with stocks from small companies, that may, by coincidence, match the return profile with a stock from a large company. Between the two algorithms, the one being closer to the SMB-factor of the standard model is the sparse label propagation algorithm, when comparing colors and scale.

Figure 22: Mean square error for the SMB-factor between label propagation and sparse label propagation.

Figure 22 demonstrates how similar the SMB-factor is between the graph signal recovery algorithms. The graph signal recovery algorithms converge (as well) to similar results, the more initial labels are fed to the stock correlation network. With respect to correlation length, the results are more similar the longer the correlation length between stocks are. As described earlier, the resulting SMB-factor has more similar results in the standard model, but the algorithms are also converging to similar results on higher correlation length. By investigating the scale on the right side, it can be concluded that the network algorithms approach more similar results to each other than the standard model with respect to the SMB-factor.
Similar to the SMB-factor, the HML-factor shows the same pattern with respect to sample size for both graph signal recovery methods. The HML-factor in both algorithms is closer to the HML-factor in the standard model, when the sample size is higher. Different to the SMB-factor, the graph signal recovery models are closer to the HML-factor of the standard model at short correlation lengths. For the HML-factor, the largest similarity to the standard model is at a correlation length of 1 year for both label propagation and sparse label propagation, and the variance of similarity is higher for the HML factor than the SMB factor in Figure 21. For the network models to be coherent with the factors from the standard model, the SMB-factor needs a longer return history in the stock correlation network, whereas HML needs a shorter return history in the stock correlation network. For the SMB-factor, the variance at the sample size-axis for sparse label propagation is lower than for label propagation. On the contrary, label propagation has less variance in similarity on the sample size-axis for the HML-factor than sparse label propagation.
Figure 24: Mean Square Error for the HML-factor between label propagation and sparse label propagation.

Figure 24 shows how similar the HML-factor is between label propagation and sparse label propagation. Both signal recovery algorithms converge, as in the previous cases, to more similar results the higher the sample size is in the stock correlation network. With respect to correlation length, the results are more similar on shorter correlation, which is consistent with previous results from Figure 23. Between the algorithms, the SMB-factor is more similar than the HML-factor between the algorithms, which can be seen from the scale on the right side between Figure 22 and Figure 24. As stated earlier, the underlying portfolios with a small number of stocks will have a larger impact on the factors, since the average return is taken on a smaller number of stocks. The realization of the uneven distribution is, therefore, a higher variance in similarity, which can be seen looking at the results from the HML-factor in Figure 23 and Figure 24 compared to the SMB-factor in Figure 21 and Figure 22.

5.3 Explanatory Power of Regression Models

The regression models are built on stocks existing longer than a year on the US stock market, producing a total of 264,000 unique models. Instead of creating a portfolio of stocks as done in most other research papers, including all stocks will induce stress testing of models and avoid selection bias. The results will be presented by averaging over each model for each of the factor construction strategies: 1) standard model, 2) label propagation algorithm and 3) sparse label propagation. Furthermore, to produce it in a presentable format, the results are presented based on the parameter values for the network approach: initial sample size and correlation length. During the regression, diagnostics will be taken to investigate whether the outcome of the
regression matches the theoretical outcome. Therefore, basic statistical tests will be conducted to investigate this matter, and if enforcing the models to pass the statistical tests, further improves the explanatory power of the model. The results are presented in two sectors: 1) all models and 2) models passing all the statistical tests. In the statistical tests, a significance level of 5 percent for rejection of the null hypothesis is used. Each table will contain at most the following information:

- **Algorithm** – The algorithm column describes which factor construction method is used, which can be either: 1) standard model, 2) label propagation or 3) sparse label propagation.

- **Correlation length** – The correlation length column refers to the total number of years that the correlation between pairs of stocks were calculated upon. Possible combinations are 1) 0 years (standard model), 2) 1 year, 3) 3 years, 4) 5 years and 5) 10 years.

- **Model count** – Model count (#) refers to the number of FF3-models created to explain the return of stocks.

- **Alpha** – Alpha is the excess return produced from the FF3 regression model (and the intersect from the multiple linear regression). The excess return presented in the table is averaged over all models.

- **$\Phi_1$** – $\Phi_1$ is the resulting slope from the FF3 regression model on the market return-factor. $\Phi_1$ presented in the table is averaged over all models.

- **$\Phi_2$** – $\Phi_2$ is the slope from the FF3 regression on the SMB-factor. $\Phi_2$ presented in the table is averaged over all models.

- **$\Phi_3$** – $\Phi_3$ is the slope from the FF3 regression on the HML-factor. $\Phi_3$ presented in the table is averaged over all models.

- **R2** – R2 is a goodness-of-fit metric between the return of the stock and the FF3 regression model. R2 presented in the table is averaged over all models.

- **R2 adjusted** – R2 adjusted is the primary metric used in this paper to measure the explanatory power of the FF3 regression model. R2 adjusted presented in the table is averaged over all models.

- **Ljung-Box ratio** – Ljung-Box ratio describes how many times the statistical test failed on the models, where it measures if the error between the return and the model is random as stated in the FF3 regression model.

- **Engle-Granger ratio** - Engle-Granger ratio describes how many times the statistical test failed on the models, where it measures if each of the three factors in the regression model is independent from the other factors or not.

- **Jarque-Bera ratio** - Jarque-Bera ratio describes how many times the statistical test failed on the models, where it measures if the distribution of the error between the model and the stock return is Gaussian.
Below are the results from the standard model and the network models, where all stocks on the market were used to fit each regression model.

Table 3: Regression results between the Fama French 3 factor model and return of stocks on the US-market between 2000-2015.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model #</th>
<th>alpha</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\phi_3$</th>
<th>R2</th>
<th>R2 Adj</th>
<th>LB-ratio</th>
<th>EG-ratio</th>
<th>JB-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>8000</td>
<td>-0.073</td>
<td>1.2</td>
<td>0.75</td>
<td>-0.012</td>
<td>0.232</td>
<td>0.181</td>
<td>0.13</td>
<td>0.91</td>
<td>0.67</td>
</tr>
<tr>
<td>LPA</td>
<td>128000</td>
<td>-0.067</td>
<td>1.2</td>
<td>0.99</td>
<td>0.36</td>
<td>0.225</td>
<td>0.173</td>
<td>0.13</td>
<td>0.91</td>
<td>0.67</td>
</tr>
<tr>
<td>SLPA</td>
<td>128000</td>
<td>-0.003</td>
<td>1.2</td>
<td>0.9</td>
<td>0.15</td>
<td>0.225</td>
<td>0.173</td>
<td>0.13</td>
<td>0.92</td>
<td>0.66</td>
</tr>
</tbody>
</table>

As can be seen in Table 3, $\phi_1$ remains the same in all three values at 1.2 (since the market value is calculated the same way independent from model used). $\phi_2$ and $\phi_3$ values have a positive slope in the network model, though the $\phi_3$ value has not as large an impact on the result as $\phi_2$ due to the low value of the slope. The alpha value is also on average closest to zero for the SLPA-algorithm, which is a good result since it incorporates the bias return of the regression model, as we are using all stocks on the market as input. The R2 values are almost the same for all three models, so there is no model category, which is significantly above the others to model return of stocks. However, the R2-values are low for all models, which means that the explanatory power for explaining the variance in the stock returns are on average low. The LB-ratio, EG-ratio and JB-ratio are almost the same for all three model categories. The failure rate of LB-ratio is at a low level, which indicates that the error between the stock return and the model is random. The EG-ratio has over a 90% failure rate, which means that in most cases the factors are not independent and the model can be shortened to a one or two factor model. The high JB-ratio indicates that the error between the model and the return from the stocks are not normally distributed, which is not desirable since the model suggests that the output would be Gaussian.
In Figure 25 can be seen the explanatory power (R2 adjusted) for both the label propagation and the sparse label propagation algorithm. The transparent surface on top, represents the explanatory power of the standard model serving as a benchmark. With respect to sample size, a high sample size results in higher explanatory power for both network models. By referencing to earlier results of similarity between the factors, high sample sizes indicated being closer to the standard model. With respect to correlation length, a smaller time window is preferred as both algorithms has a higher explanatory power at a correlation length of either 1 or 3 years. The results give an indication that the closer the network models are to the standard model, the higher amount of explanatory power will be achieved. Both label propagation and sparse label propagation are close to the explanatory power of the standard model, though none of the graph signal algorithms achieves a higher explanatory power.

By using all the stocks on the market, there was a high failure rate on the Engle-Granger test with above 90% failure rate and Jarque-Bera test with approximately 2/3 of the stocks on the market failing the test. Since the assumption in this thesis is that the Fama French model is correct and able to explain stock returns, the results produced by subtracting the stock returns with the FF3 model should be random and normally distributed (Gaussian noise). To obtain the set criteria, we enforce the null hypothesis to be accepted in the statistical tests: random output (Ljung-Box) and normally distributed (Jarque-Bera). To further ensure each factor encapturing separately the anomalies, we enforce the factors to be independent (Engle-Granger). Therefore, we will only consider stocks, that passed the three statistical tests, and investigate how it influences the explanatory power of the models.
Table 4: Regression results between the Fama French 3 factor model and return of stocks on the US-market between 2000-2015 (passing all statistical tests).

<table>
<thead>
<tr>
<th>Model</th>
<th>Model #</th>
<th>alpha</th>
<th>φ1</th>
<th>φ2</th>
<th>φ3</th>
<th>R2</th>
<th>R2 Adj.</th>
<th>LB-ratio</th>
<th>EG-ratio</th>
<th>JB-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>470</td>
<td>-0.66</td>
<td>0.95</td>
<td>1</td>
<td>-0.59</td>
<td>0.382</td>
<td>0.248</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>7300</td>
<td>-1.2</td>
<td>0.9</td>
<td>1.3</td>
<td>-0.34</td>
<td>0.388</td>
<td>0.249</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>7100</td>
<td>-1.2</td>
<td>0.89</td>
<td>1.2</td>
<td>-0.62</td>
<td>0.391</td>
<td>0.252</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

As can be seen in Table 4, the excess return is in average at a lower level of return than the market. The model with the closest alpha value is the standard model, whereas the network models are both twice less and has the same value of -1.2. In all three models the slope is positive for the SMB-factor, and the HML factor in all three models has negative slopes. In the models using graph signal recovery, the R2 and R2-adjusted have a slightly higher value than the standard model. Compared to Table 3, the average R2 and R2-adjusted values has increased for all the models, which translates to an increase in the model’s explanatory power. Under further investigation in the experiment, enforcing no cointegration between the factors increases the adjusted R2 value by almost 50% compared to earlier values Table 3.

![R2 Adjusted-value off Regression Results in Average for LPA (stat. tests passed)](image1)

![R2 Adjusted-value off Regression Results in Average for SLPA (stat. tests passed)](image2)

Figure 26: R2 adj.-value for the network models on both dimensions: sample size and correlation length (with all statistical tests passed).

As can be seen in Figure 26, the surface of the adjusted R2-value has changed when enforcing statistical tests to pass and theory to hold, compared to Figure ??.

With respect to sample size, the best performing models can be found at a lower level of sample sizes compared to earlier results. With respect to correlation length,
the best performing models are not found at a particular correlation length based on the figure. When comparing the graph signal recovery methods against the standard model, on certain parameters the graph signal recovery methods perform better than the standard model. Mainly, the increased explanatory power comes from lower sample sizes, which indicates that a model different to the standard model is better.
6 Summary and Conclusions

Attempts to understand the market and its movements have long presented a vigorous challenge. Therefore, accurately descriptive and predictive investment models have been sought after to improve the understanding of market movements in order to support different stakeholders (investors, central banks, regulators etc), in their decision making. The rise of increased computing power and consequent possibilities to enhance advanced analytics, has therefore created opportunities to capture underlying drivers of market dynamics. The main goal of this thesis was to investigate the explanatory power of the standard FF3 model as set forth by Fama French[14], for the time period 2000-2015 on the US stock market, using financial data from the Wharton Research Data Service, and to compare the results to models constructed by the graph-based semi-supervised learning framework tailored for factor models. The graph-based semi-supervised learning framework rearranges the stocks in the underlying portfolios of the Fama-French 3 factor model, where it has been shown that a higher amount of known labels evens out the distribution of stocks in the underlying portfolios. Consequently, similar FF3 models to the standard Fama French 3 factor model were created. Furthermore, the models created using the framework show that, a stock correlation network using less historical data and a higher number of known samples for graph signal recovery achieves the highest explanatory power. However, the explanatory power for the models constructed using graph-based semi-supervised learning, remained lower than the standard FF3 model. When enforcing the theoretical model to hold true using statistical filters showed that the models created by the proposed framework had higher explanatory power when the algorithms were fed lower amounts of initially known labels.

The empirical results show that the standard FF3 model has a low level of explanatory power for the selected timeframe. The low level of explanatory power in the FF3 model is probably a product of changed market dynamics, caused by e.g. the emergence of the Internet, globalization and computerized trading, all of which were not present or existed only at a smaller scale during the release of the original Fama French paper in 1992. As has been shown in previous research, existing anomalies tend to disappear quickly (such as the January effect), as they imply behavioral aspects of investor decisions. Using GB SSL shows promising results when enforcing theory to hold true. However, the low explanatory power indicates that the type of risk captured is not enclosed in the Fama French 3 factor model as is. As Fama and French argues in their 1992 paper, even though the correlation between stock returns were high during 1992 between size and book-to-market for a company, changed market dynamics has altered the nature of stock movements. Therefore, the FF3 model is unable to express a high level of explanatory power of stock returns using size and value anomalies in the market for the time period researched.

Previous research has shown that the graph topology using stock correlation networks reveal a higher level of signals in the market dynamics, and the constructed graph will result in a scale-free network. As shown in parallel by Mara[35], sparse label propagation performs better on scale-free networks, in a similar manner to how it performs for stock correlation networks excluding weaker or opposite corre-
lation values according to Tse et al[55]. Papers researching networks in the stock market[6][28][39][43][55][56][57] show that stock correlation networks contain market signals on high correlation values. In order to capture the factors driving the movements of stocks, a thorough definition of risks and anomalies contained in stocks moving in the same or opposite direction would result in a descriptive model of the market ecosystem. When that is the case, graph signal recovery would be a natural choice of making almost instant descriptive analysis of the current situation. As the empirical results in this thesis show, alternative approaches to further investigate market movements would be to either use another model reflecting the current market dynamics, or alternatively, use other metrics within the network, such as cash flow analysis, to better understand the interaction between companies on the market.

The major contribution of this thesis paved the way for a new line of research tailored for factor models. The reason the explanatory power did not exceed the explanatory power of the standard Fama French three factor model, is that the underlying mechanics driving stock returns and the market are a complex process to debunk. However, the capabilities of a combination of network analysis and graph-based semi-supervised learning are able to solve complex tasks, if the correct graph topology and algorithms are applied. The framework can also be applied in other research areas, such as biology, medicine and psychology to increase the explanatory power of existing factor models.

The research conducted in this thesis was subject to certain limitations, which, however, provide a good basis for future research. This paper has had its primary focus on reconstructing the factors in the Fama French three factor model during the time period of 2000-2015. The reconstruction was executed using a proposed framework tailored for graph-based semi-supervised learning. The graph topology is restricted to historical price returns of stocks, but other metrics of associativity such as in- and out cash flows from corporations can be utilized. When reassigning labels to stocks, the graph signal recovery algorithms label propagation and sparse label propagation were utilized, though other approaches such as label spreading or network lasso could be applied, to rearrange the stocks within the underlying portfolios of the Fama French three factor model.

In conclusion, I’ll cite the words of one of the persons who inspired me the most during my time writing this thesis: “Rearranging the stocks (in the factor construction phase) in the Fama French 3 factor model is like rearranging stones that already belong to a certain place. But go ahead and try, you might prove me wrong.” [42].
References


### A  Additional Tables

Table A1: Regression results from all parameters between the Fama French 3 factor model and return of stocks on the US-market between 2000-2015.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample Corr.</th>
<th>Model #</th>
<th>alpha</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\phi_3$</th>
<th>R2</th>
<th>R2 Adj.</th>
<th>LB-ratio</th>
<th>EG-ratio</th>
<th>JB-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>100</td>
<td>0</td>
<td>8040</td>
<td>-0.0727</td>
<td>1.21</td>
<td>0.746</td>
<td>-0.0124</td>
<td>0.232</td>
<td>0.181</td>
<td>0.126</td>
<td>0.91</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>1</td>
<td>8040</td>
<td>0.0669</td>
<td>1.12</td>
<td>0.917</td>
<td>0.416</td>
<td>0.221</td>
<td>0.168</td>
<td>0.125</td>
<td>0.932</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>3</td>
<td>8040</td>
<td>0.0415</td>
<td>1.16</td>
<td>1.04</td>
<td>0.649</td>
<td>0.221</td>
<td>0.168</td>
<td>0.126</td>
<td>0.953</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>5</td>
<td>8040</td>
<td>0.0161</td>
<td>1.17</td>
<td>1.02</td>
<td>0.411</td>
<td>0.222</td>
<td>0.17</td>
<td>0.123</td>
<td>0.909</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>10</td>
<td>8040</td>
<td>0.00689</td>
<td>1.16</td>
<td>0.951</td>
<td>0.323</td>
<td>0.218</td>
<td>0.166</td>
<td>0.124</td>
<td>0.928</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>1</td>
<td>8040</td>
<td>-0.044</td>
<td>1.18</td>
<td>1.12</td>
<td>0.541</td>
<td>0.224</td>
<td>0.172</td>
<td>0.125</td>
<td>0.908</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>3</td>
<td>8040</td>
<td>-0.097</td>
<td>1.22</td>
<td>1.16</td>
<td>0.713</td>
<td>0.225</td>
<td>0.173</td>
<td>0.125</td>
<td>0.899</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>5</td>
<td>8040</td>
<td>-0.0411</td>
<td>1.2</td>
<td>0.981</td>
<td>0.288</td>
<td>0.225</td>
<td>0.173</td>
<td>0.128</td>
<td>0.9</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>10</td>
<td>8040</td>
<td>-0.128</td>
<td>1.18</td>
<td>1.11</td>
<td>0.425</td>
<td>0.222</td>
<td>0.169</td>
<td>0.126</td>
<td>0.911</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>1</td>
<td>8040</td>
<td>-0.111</td>
<td>1.19</td>
<td>1.02</td>
<td>0.37</td>
<td>0.227</td>
<td>0.175</td>
<td>0.128</td>
<td>0.903</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>3</td>
<td>8040</td>
<td>-0.15</td>
<td>1.24</td>
<td>0.98</td>
<td>0.428</td>
<td>0.228</td>
<td>0.176</td>
<td>0.127</td>
<td>0.909</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>5</td>
<td>8040</td>
<td>-0.059</td>
<td>1.23</td>
<td>0.942</td>
<td>0.255</td>
<td>0.227</td>
<td>0.175</td>
<td>0.131</td>
<td>0.914</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>10</td>
<td>8040</td>
<td>-0.121</td>
<td>1.2</td>
<td>0.999</td>
<td>0.249</td>
<td>0.225</td>
<td>0.173</td>
<td>0.125</td>
<td>0.917</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>1</td>
<td>8040</td>
<td>-0.0898</td>
<td>1.2</td>
<td>0.92</td>
<td>0.223</td>
<td>0.229</td>
<td>0.178</td>
<td>0.126</td>
<td>0.9</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>3</td>
<td>8040</td>
<td>-0.101</td>
<td>1.21</td>
<td>0.911</td>
<td>0.258</td>
<td>0.229</td>
<td>0.177</td>
<td>0.127</td>
<td>0.917</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>5</td>
<td>8040</td>
<td>-0.0987</td>
<td>1.24</td>
<td>0.919</td>
<td>0.128</td>
<td>0.229</td>
<td>0.177</td>
<td>0.125</td>
<td>0.923</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>10</td>
<td>8040</td>
<td>-0.157</td>
<td>1.23</td>
<td>0.908</td>
<td>0.131</td>
<td>0.227</td>
<td>0.175</td>
<td>0.124</td>
<td>0.907</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>1</td>
<td>8040</td>
<td>0.218</td>
<td>1.14</td>
<td>0.724</td>
<td>0.00507</td>
<td>0.222</td>
<td>0.17</td>
<td>0.123</td>
<td>0.96</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>3</td>
<td>8040</td>
<td>0.0653</td>
<td>1.13</td>
<td>0.959</td>
<td>0.269</td>
<td>0.22</td>
<td>0.168</td>
<td>0.127</td>
<td>0.937</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>5</td>
<td>8040</td>
<td>0.144</td>
<td>1.13</td>
<td>0.79</td>
<td>0.0303</td>
<td>0.221</td>
<td>0.168</td>
<td>0.124</td>
<td>0.932</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>10</td>
<td>8040</td>
<td>-0.0428</td>
<td>1.09</td>
<td>0.895</td>
<td>0.0737</td>
<td>0.22</td>
<td>0.167</td>
<td>0.124</td>
<td>0.931</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>1</td>
<td>8040</td>
<td>0.0746</td>
<td>1.19</td>
<td>0.906</td>
<td>0.163</td>
<td>0.225</td>
<td>0.174</td>
<td>0.126</td>
<td>0.907</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>3</td>
<td>8040</td>
<td>-0.0294</td>
<td>1.21</td>
<td>0.978</td>
<td>0.389</td>
<td>0.227</td>
<td>0.175</td>
<td>0.127</td>
<td>0.902</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>5</td>
<td>8040</td>
<td>0.0197</td>
<td>1.17</td>
<td>0.922</td>
<td>0.127</td>
<td>0.224</td>
<td>0.171</td>
<td>0.127</td>
<td>0.905</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>10</td>
<td>8040</td>
<td>-0.0873</td>
<td>1.18</td>
<td>1.02</td>
<td>0.186</td>
<td>0.222</td>
<td>0.169</td>
<td>0.126</td>
<td>0.918</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>1</td>
<td>8040</td>
<td>-0.089</td>
<td>1.19</td>
<td>0.95</td>
<td>0.203</td>
<td>0.229</td>
<td>0.177</td>
<td>0.126</td>
<td>0.907</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>3</td>
<td>8040</td>
<td>-0.0857</td>
<td>1.22</td>
<td>0.929</td>
<td>0.268</td>
<td>0.228</td>
<td>0.176</td>
<td>0.128</td>
<td>0.896</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>5</td>
<td>8040</td>
<td>-0.00119</td>
<td>1.18</td>
<td>0.916</td>
<td>0.0554</td>
<td>0.226</td>
<td>0.174</td>
<td>0.129</td>
<td>0.929</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>10</td>
<td>8040</td>
<td>-0.166</td>
<td>1.19</td>
<td>1.01</td>
<td>0.123</td>
<td>0.225</td>
<td>0.173</td>
<td>0.129</td>
<td>0.906</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>1</td>
<td>8040</td>
<td>-0.0713</td>
<td>1.2</td>
<td>0.879</td>
<td>0.105</td>
<td>0.231</td>
<td>0.179</td>
<td>0.126</td>
<td>0.912</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>3</td>
<td>8040</td>
<td>-0.0692</td>
<td>1.24</td>
<td>0.891</td>
<td>0.155</td>
<td>0.23</td>
<td>0.178</td>
<td>0.128</td>
<td>0.917</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>5</td>
<td>8040</td>
<td>-0.101</td>
<td>1.25</td>
<td>0.923</td>
<td>0.0752</td>
<td>0.229</td>
<td>0.177</td>
<td>0.126</td>
<td>0.919</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>10</td>
<td>8040</td>
<td>-0.218</td>
<td>1.24</td>
<td>0.936</td>
<td>0.0927</td>
<td>0.227</td>
<td>0.175</td>
<td>0.127</td>
<td>0.906</td>
</tr>
</tbody>
</table>
Table A2: Regression results from all parameters between the Fama French 3 factor model and return of stocks on the US-market between 2000-2015 (passing all statistical tests).

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample</th>
<th>Corr.</th>
<th>Model #</th>
<th>alpha</th>
<th>φ₁</th>
<th>φ₂</th>
<th>φ₃</th>
<th>R²</th>
<th>R² Adj</th>
<th>LB-ratio</th>
<th>EG-ratio</th>
<th>JB-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>100</td>
<td>0</td>
<td>467</td>
<td>-0.658</td>
<td>0.95</td>
<td>1</td>
<td>-0.593</td>
<td>0.382</td>
<td>0.248</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>1</td>
<td>347</td>
<td>-0.663</td>
<td>0.62</td>
<td>1.02</td>
<td>-0.909</td>
<td>0.406</td>
<td>0.261</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>3</td>
<td>362</td>
<td>-0.908</td>
<td>0.753</td>
<td>1.68</td>
<td>0.282</td>
<td>0.396</td>
<td>0.244</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>5</td>
<td>492</td>
<td>-0.729</td>
<td>0.817</td>
<td>1.28</td>
<td>-0.592</td>
<td>0.397</td>
<td>0.258</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>5</td>
<td>10</td>
<td>402</td>
<td>-2.39</td>
<td>0.347</td>
<td>0.594</td>
<td>-1.12</td>
<td>0.399</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>1</td>
<td>486</td>
<td>-0.751</td>
<td>0.946</td>
<td>1.62</td>
<td>0.255</td>
<td>0.396</td>
<td>0.257</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>3</td>
<td>533</td>
<td>-0.987</td>
<td>1.16</td>
<td>1.88</td>
<td>0.606</td>
<td>0.393</td>
<td>0.26</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>5</td>
<td>530</td>
<td>-1.11</td>
<td>0.926</td>
<td>1.15</td>
<td>-0.599</td>
<td>0.382</td>
<td>0.244</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>10</td>
<td>10</td>
<td>469</td>
<td>-1.55</td>
<td>0.73</td>
<td>0.942</td>
<td>-0.908</td>
<td>0.388</td>
<td>0.246</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>1</td>
<td>517</td>
<td>-0.991</td>
<td>0.827</td>
<td>1.25</td>
<td>-0.37</td>
<td>0.391</td>
<td>0.254</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>3</td>
<td>480</td>
<td>-1.23</td>
<td>1.22</td>
<td>1.48</td>
<td>0.124</td>
<td>0.378</td>
<td>0.245</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>5</td>
<td>450</td>
<td>-1.2</td>
<td>1.21</td>
<td>1.42</td>
<td>-0.103</td>
<td>0.369</td>
<td>0.234</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>15</td>
<td>10</td>
<td>452</td>
<td>-2.27</td>
<td>0.815</td>
<td>1.18</td>
<td>-0.473</td>
<td>0.405</td>
<td>0.263</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>1</td>
<td>475</td>
<td>-0.968</td>
<td>0.911</td>
<td>1.16</td>
<td>-0.534</td>
<td>0.382</td>
<td>0.248</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>3</td>
<td>427</td>
<td>-1.14</td>
<td>1.06</td>
<td>1.3</td>
<td>-0.264</td>
<td>0.309</td>
<td>0.239</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>5</td>
<td>387</td>
<td>-1.24</td>
<td>1.03</td>
<td>1.17</td>
<td>-0.552</td>
<td>0.37</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LPA</td>
<td>20</td>
<td>10</td>
<td>503</td>
<td>-1.78</td>
<td>0.933</td>
<td>0.956</td>
<td>-0.679</td>
<td>0.387</td>
<td>0.247</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>1</td>
<td>203</td>
<td>0.35</td>
<td>0.451</td>
<td>0.21</td>
<td>-1.76</td>
<td>0.394</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>3</td>
<td>349</td>
<td>-0.847</td>
<td>0.76</td>
<td>1.37</td>
<td>-0.647</td>
<td>0.402</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>5</td>
<td>380</td>
<td>-0.915</td>
<td>0.857</td>
<td>0.823</td>
<td>-1.47</td>
<td>0.418</td>
<td>0.269</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>5</td>
<td>10</td>
<td>378</td>
<td>-2.35</td>
<td>0.343</td>
<td>0.657</td>
<td>-1.6</td>
<td>0.409</td>
<td>0.26</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>1</td>
<td>496</td>
<td>-0.939</td>
<td>0.881</td>
<td>1.58</td>
<td>-0.43</td>
<td>0.399</td>
<td>0.263</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>3</td>
<td>518</td>
<td>-0.916</td>
<td>1.05</td>
<td>1.56</td>
<td>-0.0206</td>
<td>0.388</td>
<td>0.251</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>5</td>
<td>524</td>
<td>-0.821</td>
<td>1.02</td>
<td>1.41</td>
<td>-0.68</td>
<td>0.394</td>
<td>0.258</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>10</td>
<td>10</td>
<td>449</td>
<td>-1.7</td>
<td>0.91</td>
<td>1.57</td>
<td>-0.502</td>
<td>0.396</td>
<td>0.253</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>1</td>
<td>562</td>
<td>-1.19</td>
<td>0.852</td>
<td>1.36</td>
<td>-0.432</td>
<td>0.394</td>
<td>0.256</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>3</td>
<td>540</td>
<td>-1.43</td>
<td>1.13</td>
<td>1.5</td>
<td>-0.156</td>
<td>0.388</td>
<td>0.255</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>5</td>
<td>377</td>
<td>-0.663</td>
<td>1.09</td>
<td>1.2</td>
<td>-0.727</td>
<td>0.368</td>
<td>0.232</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>15</td>
<td>10</td>
<td>500</td>
<td>-1.95</td>
<td>0.929</td>
<td>1.21</td>
<td>-0.564</td>
<td>0.389</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>1</td>
<td>468</td>
<td>-0.724</td>
<td>0.916</td>
<td>1.18</td>
<td>-0.442</td>
<td>0.387</td>
<td>0.251</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>3</td>
<td>425</td>
<td>-0.933</td>
<td>1.16</td>
<td>1.38</td>
<td>-0.246</td>
<td>0.374</td>
<td>0.245</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>5</td>
<td>412</td>
<td>-1.34</td>
<td>1.22</td>
<td>1.42</td>
<td>-0.437</td>
<td>0.374</td>
<td>0.243</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLPA</td>
<td>20</td>
<td>10</td>
<td>505</td>
<td>-1.88</td>
<td>1.03</td>
<td>1.17</td>
<td>-0.577</td>
<td>0.387</td>
<td>0.247</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>