Financial Markets and Portfolio Management (2024) 38:415–441
https://doi.org/10.1007/s11408-024-00455-4

Hidden neighbours: extracting industry momentum from stock networks

Joon Chul James Ahn[1] · **Dragos Gorduza**[2] · **Seonho Park**[3]

Accepted: 15 July 2024 / Published online: 15 October 2024

Abstract

This paper introduces an innovative method for constructing industry momentum portfolios by leveraging two stock networks: one based on stock price correlations and the other on corporate text similarity. We find that these networks capture different aspects of company relationships, motivating us to combine them and form a portfolio that exploits less visible industry momentum. Our Hidden Neighbours portfolio, analysed from 2013 to 2022, delivered an annualised return of 18.16% with a Sharpe ratio of 0.85, outperforming the S&P 500 and other traditional momentum strategies. Factor decomposition attributes returns primarily to the idiosyncratic factor α. Our study employs interdisciplinary methods, merging network analysis and Natural Language Processing (NLP) techniques for portfolio construction. Utilising advanced text embedding models, we enhance portfolio construction by integrating textual insights from corporate disclosures into stock networks. The paper offers a comprehensive strategy across diverse data and the interdisciplinary approach, uniting financial theory, network science, and NLP, advances both theory and practice of portfolio management.

Keywords Momentum · Industry momentum · Networks · Natural language processing · Portfolio management · Factor decomposition · Hidden neighbours

JEL Classification G02 · G11 · G12 · G14 · G17 · G19

✉ Joon Chul James Ahn
joonchulahn@gmail.com

Dragos Gorduza
dragos.gorduza@st-annes.ox.ac.uk

Seonho Park
seonho.park@gatech.edu

[1] Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford OX1 3JS, Oxfordshire, United Kingdom

[2] Oxford-Man Institute, University of Oxford, Oxford, United Kingdom

[3] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, United States

1 Introduction

After Jegadeesh and Titman (1993)'s seminal paper was released, researchers have found various forms of momentum profits over the years (See Wiest (2023) and references therein). One form of the momentum that was highlighted in the past literature is *industry momentum* where excess returns are generated from past price movements as investors insufficiently account for industry-wide shocks (Wiest 2023; Moskowitz and Grinblatt 1999). While various specifications of industry momentum portfolios exist, it is common to find researchers employing standard Industry Classification Schemes (ICS) such as Standard Industry Classification (SIC) and North American Industry Classification System (NAICS) (Wiest 2023). The main benefit of constructing industry momentum portfolios based on standard ICS is that it is straightforward and reproducible. However, it comes at the cost of accuracy and oversimplification of complex industry relationships between companies (Li 2022; Phillips and Ormsby 2016).

To better incorporate complex industry relationships that could not be depicted with standard ICS, researchers have been exploring the use of *networks* to understand relationships between companies and stocks. A common method of constructing stock networks is by defining the edges of a network as stock price correlation between companies (Marti et al. 2021). An alternative method is to construct stock network edges by examining the text similarity between corporate disclosures, such as 10-Ks and 10-Qs. The main objective of incorporating text data into stock networks is to capture additional dimension of information that may not have been fully incorporated in the stock price under a weaker form of the efficient market hypothesis (Fama 1970). This method involves converting corporate disclosure text into vector representations or embeddings. Nevertheless, despite the development of more sophisticated embedding models in the Natural Language Processing (NLP) literature, there has been insufficient experimentation in using these tools to build stock networks. Furthermore, while the past literature explored using either stock price correlation or text similarity networks for portfolio optimisation, there has been little attempt in using both networks for applications in portfolio management.

In this paper, we propose a novel networks-based approach to construct an industry momentum portfolio, with an aim to better capture complex industry relationships that could not be represented through standard ICS or previous network-based approaches. Our proposed industry momentum portfolio, which we refer to as *Hidden Neighbours* portfolio, is constructed using two stock networks built from stock price correlation and corporate disclosure text similarity. Our analysis will show that these two networks capture different sets of information regarding relationships between companies, prompting us to combine the two networks to capture industry momentum from less visible corporate relationships.

From 2013 to 2022, the Hidden Neighbours portfolio delivers an annualised return of 18.16% with a Sharpe ratio of 0.85, presenting superior risk-adjusted returns compared to the S&P 500 index and other well-known momentum strategies. A factor decomposition of the portfolio shows that the returns are mostly

generated through the idiosyncratic factor α. We believe the excess returns of the Hidden Neighbours portfolio stem from the identification of industry momentum between insufficiently priced-in peer companies.

The contributions of the paper can be summarised as follows:

- We suggest the use of two types of stock networks; (1) a price-based network using the stock price correlations and (2) a text-based network using the 10-K and 10-Q document embeddings. To build the text-based network, we develop and propose a novel NLP-based technique to fuse multiple document embedding techniques.
- It is analysed that those two networks have exclusive characteristics that the other does not have. As such, using both networks simultaneously, we construct Combined Network that identify peers that have strong business similarity yet have low stock price correlation.
- Using the Combined Network, a novel industry momentum portfolio, Hidden Neighbours portfolio, is proposed, delivering a Sharpe ratio of 0.85 between 2013 and 2022, while the SIC-based industry momentum benchmark delivered a Sharpe ratio of 0.55.

The remainder of this paper is organised as follows: Section 2 gives a brief literature review on industry momentum and stock network construction. Section 3 outlines our methodology in constructing stock networks using stock price correlation and corporate disclosure text data, introducing the use of modern NLP tools. In Sect. 4, we juxtapose the two networks built from different data sources and motivate the use of the Combined Network as a means to discover less visible peers of companies. In Sect. 5, we outline the portfolio construction methodology using the Combined Network and report the Hidden Neighbours portfolio's performance between 2013 and 2022. In Sect. 6, we discuss the distinct industry peers generated through our networks-based approach, and other considerations relevant to the Hidden Neighbours portfolio, such as maximum drawdown and trading costs. Finally, we provide the concluding remarks in Sect. 7.

2 Background and related works

2.1 Momentum in stock returns

Momentum is an economic anomaly where buying stocks with positive past returns and selling the negative yielding ones deliver positive returns (Wiest 2023). This market anomaly began to be discussed extensively in the academic literature after Jegadeesh and Titman (1993) paper, which showed that one can generate excess profits simply by looking at past returns. After a series of different investigations on the source of momentum profits, it is now widely accepted in the academic community that multiple forms of momentum profits exist across different time periods and asset classes (Wiest 2023).

One explanation for the prevalence of momentum profits is investor under/over-reaction to information (Daniel et al. 1998). As investors have innate biases, such as self-attribution and overconfidence (Chui et al. 2010), they fail to fully incorporate new information, leading to lagged positive or negative returns. Such inefficiencies are more pronounced in smaller and lower coverage markets, leading to larger momentum profits (Hong et al. 2000).

2.2 Industry momentum

Industry momentum is one form of momentum deeply researched in the literature, where there is a continuation of excess returns among companies in the same industry (Wiest 2023). The extensive review of industry momentum and momentum-based approaches is provided in Wiest (2023). To the best of the authors' knowledge, Moskowitz and Grinblatt (1999) is the first work that investigates industry momentum by grouping companies based on their first 2 digit SIC code. They constructed a long-short portfolio that holds companies in the top 3 highest-returning industries and shorts the bottom 3 lowest-returning industries. However, unlike the momentum profits documented in Jegadeesh and Titman (1993), industry momentum profits showed the highest returns with 1 month holding period and exhibited a faster decline in excess profit with longer holding periods. A long-only sector fund implementation of industry momentum strategy was unable to outperform the S&P 500 index on an risk-adjusted basis between 1989 and 1999 (O'Neal 2000).

Similar to Moskowitz and Grinblatt (1999), many researchers have utilised standard Industry Classification Schemes (ICS), such as SIC codes, to group companies into their assigned industries before extracting industry momentum (Li 2022; Grobys and Kolari 2020; Behr et al. 2012). Ease of implementation and repeatability of ICS-based groupings is the main advantage of using this method. However, standard ICS tends to oversimplify complex industry relationships between companies as it forces a company to be identified by a single industry code. Furthermore, the magnitude of excess returns from industry momentum profit tends to vary depending on the choice of ICS, creating confusion among practitioners who need to make an arbitrary choice on the specific ICS to use (Li 2022).

To overcome the limitations of ICS-based groupings, Hoberg and Phillips (2016) used product descriptions of 10-K documents to construct frequency-based word vectors and group companies into industries based on their product similarity. The industry momentum portfolio constructed on their text-based classification was able to generate longer and larger excess returns compared to SIC-based industry momentum portfolio (Hoberg and Phillips 2018). In this work, the authors claimed that the longer and larger industry momentum profits stem from the ability to identify less visible industry relationships through text-based classification (Hoberg and Phillips 2018).

2.3 Network of stocks

To better depict complex industry and business relationships, researchers have explored the use of networks to analyse relationships between different companies

and their stocks (Marti et al. 2021). While industry classification forces one company to be identified through a single code within an overall hierarchy, a network approach can be adopted to capture more complex business relationships between stocks (Hoberg and Phillips 2016). The most common method to construct stock networks is to use return correlation data (Marti et al. 2021). Stocks are represented as nodes and their total return correlations are used to define the edge weights, resulting in a weighted network of stocks (Mantegna 1999). After defining the nodes and the edges, it is common to construct a Minimum Spanning Tree (MST) and examine the clusters of companies formed (Marti et al. 2021).

While such stock price-based networks have shown to be good risk management tools (Lee and Nobi 2018), they may not be suitable for industry representation. An MST structure forces an acyclic structure, which may be too restraining on representing complex real-life industry relationships. Furthermore, the presence of spurious correlation among stocks could be a source of error, leading to connections between nodes even when no real economic relationships are present.

Thus, a text-based network that analyses corporate disclosure text, such as the one suggested in Hoberg and Phillips (2016), could be an alternative consideration to depict industry relationships as it can account for product description, industry jargon, common business risk, and so forth. However, it imposes additional complexity as researchers need to choose an appropriate methodology to convert text data to vector embedding representation. Notably, Hoberg and Phillips (2016) converted 10-K product descriptions into vector representation using a word-count-based approach. Although a word-count-based approach is interpretable, it is unable to fully account for the overall syntax and it is not robust to changes in product description over time. Applying a machine learning-based document embedding model, Adosoglou et al. (2022) used the Doc2Vec (Le and Mikolov 2014) model to construct text-based networks, uncovering business text similarity beyond product similarity. Using more state-of-the-art embedding models allow researchers to extract more information from corporate disclosure text, leading to more informative text-based networks. However, trade-offs remain as these models tend to be less interpretable, potentially leading to hidden biases.

2.4 Portfolio construction leveraging text information

With an increase in computational power and online content, there has been an exponential increase in use of text data for financial analysis and portfolio construction (Loughran and McDonald 2020). Among the wide range of text data used in prior research (Loughran and McDonald 2020), one common source of text data is corporate disclosures filed to the U.S. Security Exchange Commission (SEC), such as 10-K and 10-Q disclosures (henceforth referred to as 10-X disclosures). 10-X disclosures contain valuable information about companies, such as their general business description and key risk considerations (Dyer et al. 2017). Hypothesising that the market insufficiently prices in text data in 10-X disclosures, Cohen et al. (2020) constructed a market neutral portfolio that holds a long position on companies that exhibit the least change in their 10-X disclosures while shorting companies that

exhibit the large changes. The "Lazy Prices" (Cohen et al. 2020) portfolio generated the monthly α of 188 basis points, suggesting that there is a significant amount of information embedded in 10-X disclosures that could be exploited for constructing the profitable portfolio.

Extending the "Lazy Prices" portfolio construction methodology, Adosoglou et al. (2022) proposed the "Lazy Network" portfolio, which incorporates a network-based methodology to analyse the correlations between companies that least changed their corporate disclosures over time. In Adosoglou et al. (2022), the long-only equal-weighted portfolio of 50 companies that least changed their corporate disclosures, selected through various centrality measures within a network, generated the monthly α of ranged between 51 and 96 basis points. Given that the "Lazy Network" portfolio only holds long positions, it can be said that the extent of the abnormal return is similar to that of the "Lazy Price" portfolio in Cohen et al. (2020).

When using the corporate disclosure text data to construct portfolios of stocks, the choice of language model can influence the size of excess returns. Adosoglou et al. (2021) compared the excess return of financial portfolios constructed based on text similarity using three different language models, Word2Vec (Mikolov et al. 2013) and two Doc2Vec (Le and Mikolov 2014) implementations; PV-DM and PV-DBOW. The largest excess return was generated by the portfolios that utilised the PV-DM Doc2Vec model, implying that language models which can account for the order and semantics of words tend to perform better when constructing portfolios that utilise text similarity.

3 Constructing text-based and price-based networks

3.1 Price-based network

In the networks that the paper constructs, the nodes represent the companies in the S&P500 at the end of the calendar year. The edges of the *price-based network* are the stock price correlation between two companies. The historical price information we used in this paper was retrieved from Center for Research in Security Prices.[1] Given the stock price P_i of a stock i, the daily return Y_i can be defined as follows:

$$Y_i = lnP_i(t) - lnP_i(t - \Delta t), \tag{1}$$

where $\Delta t = 1$. Then, the Pearson product-moment correlation ρ_{ij} between two stocks i and j is calculated as follows (Birch et al. 2016):

$$\rho_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}}, \tag{2}$$

[1] US Stock Database © 1962 Center for Research in Security Prices, LLC, An Affiliate of the University of Chicago Booth School of Business.

where the $\langle \cdot \rangle$ operation represents an average over the calendar year. Most stock price-based networks in the literature are constructed using stock price return correlation formulated above (Marti et al. 2021), which can be obtained readily.

3.2 Text-based network

Besides the price-based network, the text-based network is also introduced in this paper. The text-based network is constructed with the same nodes as the price-based network, but the edges of it are defined as the cosine similarity between the vector representations of text data for companies. As in Adosoglou et al. (2022), we also adopted the corporate disclosures, 10-K and 10-Q (10-X in general), as the text information we analyse.

3.2.1 Text data collection

The historical 10-X disclosures are gathered from the Edgar Crawler developed in Loukas et al. (2021) and The Notre Dame Software Repository for Accounting and Finance.[2] To prevent any look-ahead bias, we strictly collected disclosures within the calendar year instead of the fiscal year.

3.2.2 NLP models

For each calendar year, we aim at constructing a text-based network. Thus, it is necessary to define the edges of it, which are not mutable during a year. Similar to Hoberg and Phillips (2016), we determine the edges of the text-based network as the cosine similarity of yearly document vector representation or embedding. To do so, we adopted two NLP models specifically. Furthermore, we also suggest a way of consolidating the outputs of 2 NLP models seamlessly. We specify the training methodologies below.

Doc2Vec

Doc2Vec (Le and Mikolov 2014) is a widely used bag-of-words document representation method, which is trained on a collection of words with a paragraph ID for each document. The addition of paragraph ID extends the Word2Vec Mikolov et al. (2013) embedding method to the Doc2Vec giving more flexibility in representing a bag-of-words contained within a paragraph. We train the Doc2Vec model with the hyperparameter settings suggested in Adosoglou et al. (2022). These settings are reasonable because the 10-X data we have used in this paper is almost identical to the data in their work. Details of the training methods are specified in Appendix A.

[2] https://sraf.nd.edu/data/stage-one-10-x-parse-data/.

FinBERT

While the Doc2Vec model has been successfully deployed for creating a text-based network in the previous works (Jeon et al. 2017; Adosoglou et al. 2022), this model has clear limitations that it is a bags-of-words approach. Doc2Vec is unable to represent the entire syntax as it does not account for the order of words in a given text. Thus, we seek to further improve the text-based network by adopting a more recent NLP model that can better account for the overall syntax. Specifically, we investigated the use of FinBERT (Huang et al. 2022). FinBERT is a Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al. 2019) that is pretrained on corporate disclosure and financial text data, making it suitable for use in financial applications. By using FinBERT, we want to create more complex document vector embedding that can better account for nuanced similarities between corporate disclosures from different companies.

To create document-level embedding using FinBERT, we first fine-tune the FinBERT model using SimCSE (Gao et al. 2021) contrastive learning method. To circumvent the labelling process for training, SimCSE contrastive learning, a self-supervised learning method, utilises the distance between the vector representations for sentences from the text instance. Specifically, it tries to minimise the distance between the vector representations from the same sentence while trying to maximise that from different sentences. For generalising the representation, the dropout layer (Srivastava et al. 2014) is also attached to the penultimate layer of the implemented FinBERT model. While it may be also possible to fine-tune the FinBERT model using ICS such as SIC or NAICS as our labelled data, this will simply reinforce the bias and inaccuracy present in assigned industry codes, and diminish our efforts to understand more nuanced relationships. The overall process of fine-tuning using SimCSE is illustrated in Fig. 1.

After fine-tuning, we use the FinBERT model to yield vector representation for sentences in the 10-X document. Then, the sentence embeddings are averaged out to generate our final document embedding with a dimension of 784. Similar to Doc2Vec, we average the 10-X disclosures to create an annual representation. This process is summarised in Fig. 2. Further details on the training method are elaborated in Appendix A.

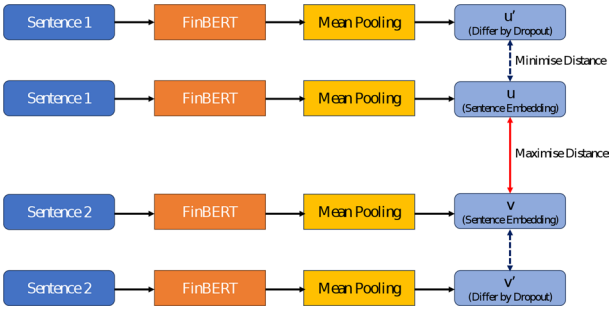

Fig. 1 The FinBERT fine-tuning process using SimCSE contrastive learning

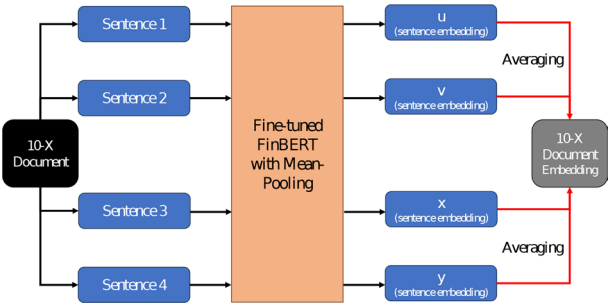

Fig. 2 Document embedding generation using the FinBERT model after fine-tuning

Combining Doc2Vec and FinBERT

Based on the two NLP models suggested above, we are able to calculate cosine similarity between corporate disclosures of different companies to complete the text-based networks. For each NLP model, we calculate cosine similarity (*CS*) between corporate disclosures by:

$$CS = \frac{\Sigma_i x_i y_i}{\sqrt{\Sigma_i x_i^2}\sqrt{\Sigma_i y_i^2}}, \tag{3}$$

where x_i and y_i are the ith components of two different annual document representations of the same size.

From the calculation above, we get two sets of cosine similarities between companies; one constructed on the Doc2Vec model and the other on the FinBERT model. We finalise the text-based network edge construction by averaging the cosine similarity between these two sets. We used this combined methodology as it delivered us the highest risk-adjusted return, which will be discussed later in Sect. 6.4.

3.3 Network backboning

Network construction often involves a backboning process, which refers to the removal of insignificant edges to simplify the structure of the network. Network backboning is essential for both text-based and price-based networks, as they are constructed based on document embedding cosine similarity and stock price correlation. Given that most companies have nonzero text similarity and stock price correlation, without backboning, each node in the text-based and price-based network will be fully connected to other nodes. Not only will visual identification suffer in this fully-connected structure, it is a poor representation of real-life industry relationships.

To backbone stock networks, the global thresholding method is often used (Marti et al. 2021), which removes any edge weight below a certain cut-off parameter. However, as shown in the left panel of Fig. 3, the edge weight distributions of price-based and text-based networks differ significantly. Applying the global thresholding

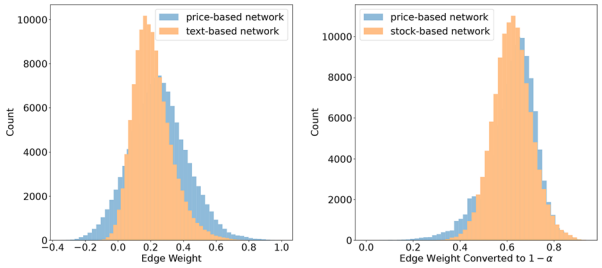

Fig. 3 Distributions of edge weights before (left) and after (right) applying the normalisation using the disparity filter

method is inappropriate in this case as it leads to a biased representation of either one of the networks.

Therefore, instead of a global thresholding method, we backboned both the price-based network and text-based network in a consistent manner using disparity filter (Serrano et al. 2009). The disparity filter method attempts to locally determine the statistically significant edges of a node by assuming a null model where the normalised weights of the degree of the node k follow a random assignment from a uniform distribution. The disparity filter converts the edge weight into a $1-\alpha$ test statistic and drops any edge with an $1-\alpha$ statistic lower than a pre-set level, where α is the Type 1 error. The $1-\alpha$ statistic is calculated as follows:

$$1-\alpha = (k-1)\int_0^{p_{ij}} (1-x)^{(k-2)}dx = 1-\left(1-p_{ij}\right)^{k-1}, \tag{4}$$

where p_{ij} is the normalised weight between nodes i and j, i.e. $p_{ij} = \frac{w_{ij}}{\Sigma_i w_i}$, and k is the degree of the node under consideration.

After applying the disparity filter, we treat $1-\alpha$ from Eq. (4) as an edge weight. Figure 3 shows the change in the distribution of edge weight before and after conversion into $1-\alpha$ statistics. One can observe that after the conversion, the distribution of text-based and price-based network edges is a lot more similar to each other, giving us better grounds for setting a common $1-\alpha$ cut-off to backbone both networks.

After normalising the weights of two networks using the disparity filter, it is recommended to set the cut-off for backboning the networks as high as possible, as this would remove any statistically insignificant edges from both networks. However, when the cut-off value is set too high, it may result in loss of edge weight and nodes, leading to an oversimplified sparse network, or network isolation. Since the final goal of this paper is portfolio optimisation, the loss of nodes from the backboning process will not be ideal. Therefore, we chose 0.7 as our cut-off value. We found this value to be sufficiently high to remove statistically insignificant edges, while not causing any loss of nodes in our experiments. Further details and considerations are discussed in Appendix B.

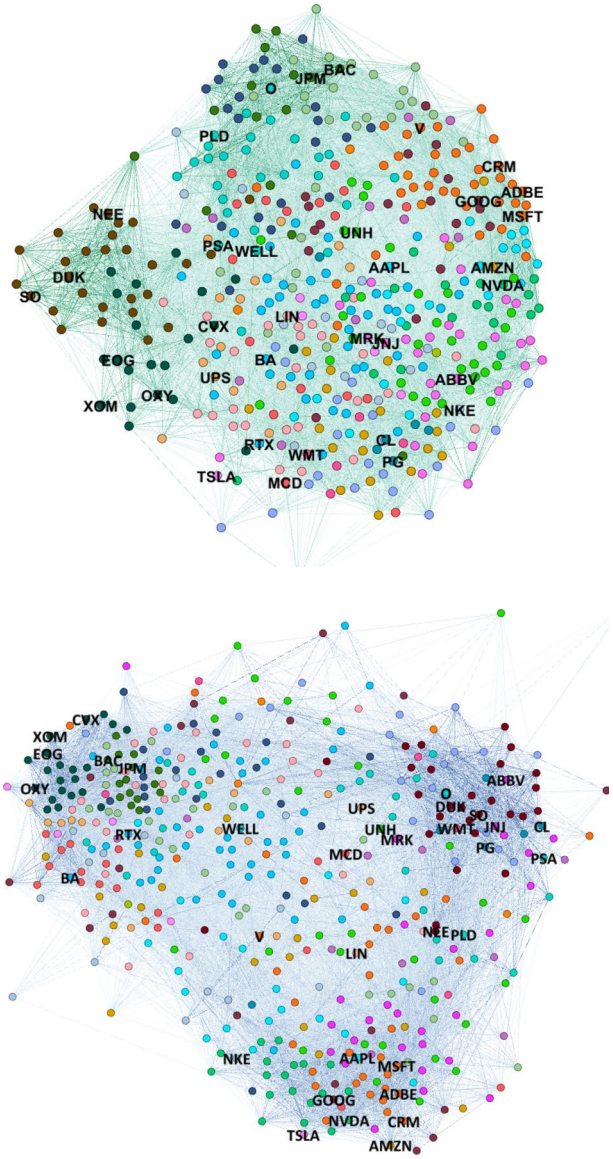

Fig. 4 Text-based network (top) and price-based network (bottom) of S&P 500 companies based on 2021 data, where each node is colour coded by its S&P Capital IQ primary industry classification

4 Constructing combined network

4.1 Network analysis

In this section, we conduct network analysis on text-based and price-based networks to understand if the two networks capture different sets of information regarding relationships between companies. This network analysis, in turn, motivates the use of the Combined Network to identify the less visible relationships between companies.

Visualisation

Based on the methodologies outlined above, we created the text-based and price-based networks, and visualised them using Gephi (Bastian et al. 2009) as shown in Fig. 4.[3] From this visualisation, one can see that text-based network display clusters of companies that align well with S&P Capital IQ primary industry classification. Notably, we see clusters of Software and Services companies (orange) at the right-hand side of the network, situated opposite from Utilities companies (brown) at the left-hand side of the network.

However, for the price-based network, clusters are more aligned with their risk-return profile rather than their industry classification. The bottom section of the price-based network consists mostly of growth stocks such as Tesla (TSLA), Amazon (AMZN), and Google (GOOG), indicating a group of more volatile stocks. Conversely, the top right-hand corner displays companies with lower beta, such as Walmart (WMT), Duke Energy (DUK), and Realty Income (O).

A notable example we observe is the location of TSLA within both networks. In the text-based network, where the description of business and core operations are considered, TSLA is situated far from Software Services companies as its core business of electric car sales is different from software services. However, in the price-based network, the TSLA node is situated near Software Services companies. This could be due to the fact that both TSLA and many Software Services companies are considered as growth stocks with a high-risk high-return profile, resulting in a high stock price correlation.

Based on visual inspection, it can be inferred that the two networks are different from each other; the text-based network displays relationships more aligned with business activity, whereas the price-based network displays relationships based on the risk-return profile of stocks. We conjecture that relationships in text-based networks are more driven by fundamental business similarity, while relationships in price-based networks are more strongly influenced by market-related factors and investor perceptions.

[3] The detailed visualisation can be found in https://sites.google.com/view/hn-network-vis.

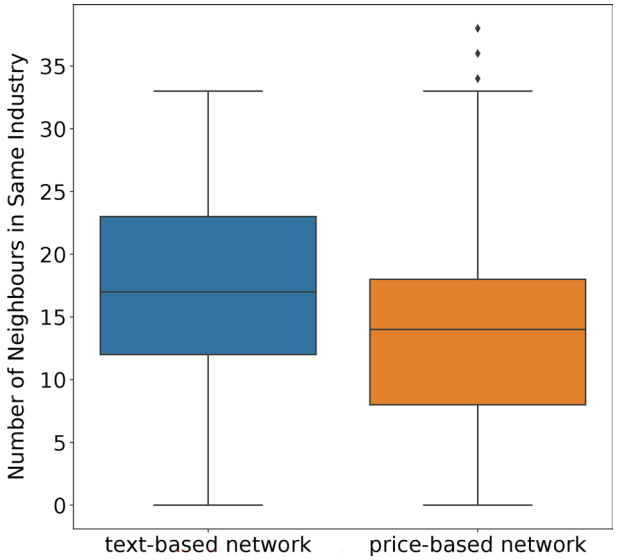

Fig. 5 The total number of neighbours belonging to the same industry classification as the respective node

Table 1 Structural similarity of text-based and price-based networks across 2012–2021, measured on an annual basis

Period	Text-based network	Price-based network
2012–2013	0.292	0.173
2013–2014	0.279	0.214
2014–2015	0.276	0.229
2015–2016	0.251	0.195
2016–2017	0.276	0.191
2017–2018	0.268	0.202
2018–2019	0.230	0.206
2019–2020	0.241	0.217
2020–2021	0.237	0.237

Higher number indicates lower structural similarity

Industry classification of nearest neighbours

By analysing the nearest neighbours of nodes in both networks, we want to further analyse the clusters of companies formed in each network. Specifically, we want to understand which network forms clusters more akin to a traditional ICS by counting the number of neighbours which share the same ICS with the respective node.

Figure 5 shows the distribution of the number of nearest neighbours with the same industry classification as the respective node, based on S&P Capital IQ Primary Industry Classification. It can be observed that the nearest neighbours of text-based network are more likely to be in the same industry classification of the respective node than the price-based network. The findings are congruent with our visual inspection, where standard ICS clusters were more visible under the text-based network. The analysis of neighbours shows that while both networks capture economic relationships between companies, the text-based network is closer to representing relationships more akin to traditional ICS. This analysis can prove useful to portfolio builders looking to ascertain the information content of the presented networks relative to an industry baseline.

Network dynamics across time

Stock networks evolve across time with new information. Given that the data sources of text-based and price-based networks differ, we want to analyse the change in network structure across time. Using graph similarity Faizliev et al. (2019), which is a weighted average of Hamming Distance and change in PageRank centrality, we measured the annual change in the structural similarity of the text-based and price-based networks. The graph similarity measure ranges from 0 to 1, with 0 indicating that the network is identical to its previous year's structure and 1 indicating complete dissimilarity.

Table 1 shows the calculated graph similarity values for 1 year periods between 2012 and 2021, tracking the annual change in the network structures. Overall, price-based network shows higher similarity to its prior year's network, compared to the text-based network, with a relatively lower similarity value across each observation period. While one may expect the price-based network to show greater structural change over time, as it is built on stock price correlation which is dependent on investor activity, our analysis shows otherwise. The text-based network shows a greater degree of structural change across all years, compared to the price-based network.

The smaller change in network structure in the price-based network could be due to slow adaption by investors and their sticky behaviour in established industry relationships. As documented in the past literature, investors can show limited attention and stickiness in existing industry relationships (Kimura and Nakagawa 2022). Even during periods of market stress, there was no evidence of investor herding behaviour or changes in their allocations in sector ETFs (Gleason et al. 2004). While it may be early to conclude any lead-lag relationships, our analysis shows that the two networks differ in terms of their structural change across time.

4.2 Combined network

Our network analysis above shows that while the two networks both depict relationships between S&P 500 companies, each network captures different sets of

information. Notably, we find evidence that the price-based network is more market-oriented, while the text-based network is more closely aligned with business fundamentals. While the previous works have focused on utilising either text-based or price-based network alone, recognising that these two networks capture different information, we are motivated to combine the two networks for better optimisation.

We propose the use of the Combined Network, which subtracts the edge weight of the price-based network from the text-based network. This could be done by subtracting the adjacency matrix of the two networks:

$$C_{ij} = A_{ij}^{text} - A_{ij}^{price} \tag{5}$$

where i, j refer to the S&P500 nodes, A^{text}, A^{price} refer to the adjacency matrices of the text-based and price-based network, respectively, and C_{ij} is the adjacency matrix of the Combined Network. Through this operation, we seek to discount industry relationships, which are already "priced-in" by investors in the market, reflected in the price-based network on the textual relationships.

Thus, with the Combined Network, we wish to observe industry momentum on companies that have strong business similarities, yet have low stock price correlation with each other. We refer to such companies as Hidden Neighbours, peers of companies in which their relationship is less visible and insufficiently priced by the market.

5 Hidden neighbours industry momentum portfolio

5.1 Hidden neighbours portfolio

Using the Combined Network, we seek to construct a long-only industry momentum portfolio with superior risk-adjusted returns compared to different benchmarks.

Our proposed industry momentum portfolio is constructed in the following manner. First, we define momentum stocks as stocks with top 30th percentile total returns in the past $12-1$ months, with 1 month skipping in the look-back period. Secondly, we rank each node in the Combined Network by the average of edge weights connected to momentum stocks, i.e. the highest rank is assigned to nodes with the highest average Combined Network edge weights with momentum stocks. Lastly, we select the top 50 stocks among this ranked list to be included in our portfolio on an equal-weighted basis, and hold them for 12 months.

We name this industry momentum portfolio the Hidden Neighbours portfolio, as it seeks to choose stocks that have strong business similarity yet with low stock price correlation with momentum stocks. We believe this methodology provides us with an avenue to capture industry momentum among less visible peer group companies, where the market is under-pricing their business similarity. Examples of Hidden Neighbours peers are discussed in Sect. 6.1.

Table 2 Returns (%) and Sharpe ratio of Hidden Neighbours and other benchmark strategies between 2013 and 2022

Strategy	Sharpe ratio	Annualised returns	Cumulative returns
Hidden Neighbours	0.85	18.16	457
Standard Momentum	0.73	17.06	393
SIC Industry Momentum	0.55	13.72	290
S&P 500	0.57	12.02	268

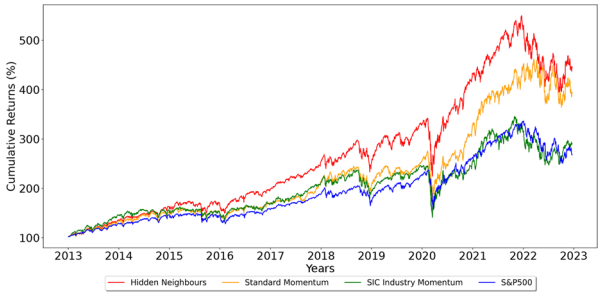

Fig. 6 Cumulative returns of the Hidden Neighbours portfolio against benchmark other momentum strategies

5.2 Benchmarks

We gauge the relative performance of the Hidden Neighbours portfolio by comparing it against three benchmarks:[4]

- *Standard Momentum* (Jegadeesh and Titman 1993) Standard Momentum benchmark is computed by selecting the top 50 stocks in the S&P500 with the highest total returns, with a $6-1$ look-back period. We hold this portfolio for 6 months, resulting in $J = 6, K = 6$ equal-weighted momentum portfolio with 1 month skipping.
- *SIC Industry Momentum* (Moskowitz and Grinblatt 1999) Specifically, we first identify momentum stocks as stocks with top 30^{th} percentile total returns. Then, industry momentum is extracted using the first 2 digits of a company's SIC code, which appear among the momentum stocks. Finally, all S&P 500 companies with the respective momentum-experiencing SIC codes are selected, and we equally weight the top 50 stocks with the highest total return in a $6-1$ look-back period. The holding period is set to 6 months.
- *S&P 500 Index* We compare our returns to the market portfolio, which was inferred from the total return of the S&P 500 index.

[4] Our benchmarks deviate from the exact specification suggested by each paper as they have adjustments made for fair comparison. More details on benchmark construction can be found in Appendix C.

5.3 Sharpe ratio and cumulative returns

We report the Sharpe ratio and returns of the Hidden Neighbours portfolio between 2013 and 2022 along with the benchmark performances. As shown in Table 2, the Hidden Neighbours portfolio strongly outperforms other benchmark strategies. Not only was it able to deliver higher risk-adjusted returns, but also it created the most wealth with the highest cumulative returns. Conversely, while the SIC Industry Momentum benchmark was able to generate wealth, it came at the cost of additional volatility, resulting in a lower Sharpe ratio than the S&P 500 index (Fig. 6).

Despite a relatively long holding period of 12 months, the Hidden Neighbours portfolio was able to sustain its outperformance during the observation period. The persistent outperformance is similar to the text-based industry momentum portfolio suggested in Hoberg and Phillips (2018), where the identification of less visible peers led to longer return shocks up to 12 months. On the other hand, industry momentum among SIC-based classification tends to reverse its excess profits within a shorter period of time (Hoberg and Phillips 2018; Moskowitz and Grinblatt 1999), and our benchmark results in Table 2 support this prior finding. The persistent outperformance of the Hidden Neighbours portfolio also makes the strategy immune to Grundy and Martin (2001) criticism on the influence of short-term autocovariance in industry momentum portfolios.

5.4 Factor decomposition

We also decompose the return profile of Hidden Neighbours portfolio using the Carhart's Four Factor model (Carhart 1997), specified as follows:

$$R_t = \alpha + \beta_{market}(R_m - R_f) + \beta_{HML}HML_t + \beta_{SMB}SMB_t + \beta_{UMD}UMD_t + \epsilon_t. \quad (6)$$

The Carhart's Four Factor Model is used since it includes the Standard Momentum factor variable *UMD*. By analysing β_{UMD}, we can better understand if the portfolio's return is different from Standard Momentum measured by simple past returns. We are also interested in Carhart's α, which helps us understand if the Hidden Neighbours returns are driven by idiosyncratic security selection that could not be explained by standard style factors.

Table 3 The daily returns of Hidden Neighbours and benchmark momentum portfolios using the Carhart's Four Factor model

Cahart's factors	Hidden Neighbours	SIC Industry Momentum	Standard Momentum
α	0.06 (0.024) **	0.05 (0.027)	0.06 (0.030)
$R_m - R_f$	0.10 (0.055)	0.10 (0.064)	0.11 (0.057) **
SMB	0.11 (0.064)	0.12 (0.072)	0.12 (0.067)
HML	0.01 (0.054)	−0.01 (0.063)	0.12 (0.067)
UMD	0.04 (0.030)	0.05 (0.036)	0.06 (0.026) **

The numbers in parenthesis are standard errors

** indicates statistical significance at 5% significance level

Table 4 Five example peer companies generated from respective methodologies

SIC	NAICS	Price-based network	Text-based network	Combined network
State Farm	AIG	Bank of America	W.R. Berkley	The Hartford
Travelers	Progressive	MetLife	Cincinnati Financial	DR Horton
The Hartford	Travelers	Prudential	Allstate	CVS Health
MetLife	Liberty Mutual	J.P. Morgan	AIG	Cardinal Health
Prudential	Loew's	Loew's	Norfolk Southern	J.M. Smucker

For SIC and NAICS, similar peers were chosen based on primary classification. For price-based, text-based, and Combined Network, peers with the top 5 highest edge weights were chosen

Table 3 shows the regression results of fitting the portfolio's daily returns against the Carhart's Four Factor model (Carhart 1997), along with other momentum benchmarks. The Hidden Neighbours' returns display a statistically significance α at 5% significance level, showing that the excess returns are driven by idiosyncratic security selection rather than standard style factors. This is in accordance with findings in Lewellen et al. (2010), which showed that cross-section industry returns are poorly captured by standard style factor models.

Similarly, the SIC Industry Momentum shows low statistical significance with standard factors in Carhart (1997) on a 5% significance level. However, the strategy fails to generate statistically significant α. Given that the SIC Industry Momentum benchmark has a 6-month holding period, the lack of α could be due to the relatively fast decay of industry momentum shocks when identified through publicly visible SIC classifications (Hoberg and Phillips 2018).

Size of abnormal returns

For many researchers, the size of abnormal returns, or Cahart's α, is an important consideration when reviewing a portfolio. A portfolio with high α suggests that the return is mainly driven by security selection unique to the underlying portfolio construction methodology. The annualised α of our Hidden Neighbours portfolio is approximately 15.0%. Compared to the "Lazy Network" portfolio (Adosoglou et al. 2022), which had an annualised α of 11.5%, the Hidden Neighbours portfolio delivered a higher α by approximately 3.5%. Compared to the "Lazy Prices" portfolio (Cohen et al. 2020), which had an annualised α of 22.6%, the Hidden Neighbours portfolio delivered a smaller α. However, one must take into account that the "Lazy Prices" portfolio is constructed as a long-short portfolio, which means that the α figure constitutes of abnormal returns from both the long-end and short-end of security selection. Given that our Hidden Neighbours portfolio is constructed as a long-only portfolio, it may be inappropriate to directly compare the α between "Lazy Prices" (Cohen et al. 2020) and our proposed Hidden Neighbours portfolio.

While the size of abnormal return should not be the sole barometer of successful portfolio construction, compared to similar portfolios suggested in prior literature, we believe that the Hidden Neighbours portfolio delivered a statistically

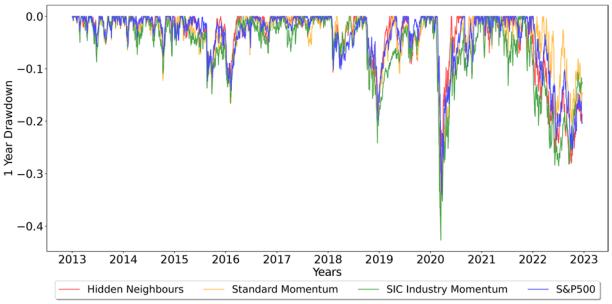

Fig. 7 Fifty-two-week maximum drawdown of Hidden Neighbours portfolio and other benchmarks

Table 5 Fifty-two-week maximum drawdown (%) on two different bear market periods

Strategy	2020	2022
Hidden Neighbours	−34.35	−28.20
Standard Momentum	−39.37	−21.20
SIC Industry Momentum	−42.68	−28.51
S&P 500	−33.92	−25.43

significant α with a meaningful size. Other elements, such as trading cost and maximum drawdown, are important qualities to consider in a portfolio, and will be discussed in Sect. 6.

6 Further discussion

Our results showed that the Hidden Neighbours portfolio delivered higher Sharpe ratios than our benchmark considerations, with a statistically significant Cahart's α. Below, we investigate other considerations relevant to the Hidden Neighbours portfolio.

6.1 Identifying hidden neighbours with combined network

As the Hidden Neighbours portfolio captures nodes in which its neighbours are experiencing strong momentum in the Combined Network, we believe the source of Hidden Neighbour's α stems from the identification of less visible industry peers in the Combined Network. Table 4 showcases five peer companies of Berkshire Hathaway generated through standard ICS and other network-based approaches. Berkshire Hathaway is an insurance conglomerate that owns companies across numerous sectors. While the peers generated from standard ICS methodologies revolve around Berkshire Hathaway's core GEICO insurance business, the Combined Network is better able to account for the conglomerate's non-core businesses, spanning across railroad, healthcare, food & beverage, etc. The Combined Network discounts

relationships already established by price-based network from the text-based network, leading to the discovery of peers closer to Berkshire Hathaway's non-insurance periphery businesses.

While it may be argued that the peers generated from the Combined Network have lower business similarity compared to standard ICS or price-based and text-based networks approach, extracting industry momentum from less visible peers in the Combined Network could translate to larger excess momentum profits as investors show lower attentiveness to these peers. The outperformance of our Hidden Neighbours portfolio against the SIC Industry Momentum benchmark supports the above hypothesis.

6.2 Maximum drawdown

One criticism of momentum portfolios is that they are susceptible to momentum crashes, suffering from large drawdowns where most of the excess profits are reversed (Barroso and Santa-Clara 2015). We measure the Hidden Neighbours portfolio's maximum drawdown from its 52 weeks high to understand how susceptible the strategy is to momentum crashes and market dislocations.

From Fig. 7 and Table 5, it can be seen that the maximum drawdown of Hidden Neighbours is similar to the S&P 500 and other benchmark strategies. The 28% maximum drawdown in 2022 is discouraging, and other strategies such as realised volatility targeting (Barroso and Santa-Clara 2015) could be implemented to reduce the maximum drawdown (Table 6).

6.3 Trading cost and portfolio turnover

The Hidden Neighbours portfolio is robust to erosion of returns from trading costs. While the benchmark momentum strategies are rebalanced every 6 months, the Hidden Neighbours portfolio is rebalanced every 12 months. Thus, all else constant, the trading cost of the Hidden Neighbours portfolio will be approximately half of the benchmark momentum strategies.

Furthermore, the Hidden Neighbours portfolio shows low annual portfolio turnovers with a median annual turnover of 38%. This is significantly lower than the SIC Industry Momentum benchmark, which showed a 75% median annual turnover, as well as many other momentum strategies introduced in the literature with a typical portfolio turnover close to 100% (Li et al. 2009; Baltas and Kosowski 2012).

Table 6 The Sharpe ratio of Hidden Neighbours portfolio using different embedding methodologies for constructing text-based network

Embedding method	Sharpe ratio
Doc2Vec	0.75
FinBERT	0.63
FinBERT w/o fine-tuning	0.56
Doc2Vec+FinBERT (Proposed)	0.85

Overall, we believe that the Hidden Neighbours portfolio is advantageous to other momentum strategies when trading cost is taken into consideration.

Using the past trading cost of S&P 500 constituents (Frazzini et al. 2018), we conservatively estimate the annualised trading cost of the Hidden Neighbours portfolio to be approximately 400 basis points, which is almost twice larger than the trading cost assumed in Moskowitz and Grinblatt (1999), Jegadeesh and Titman (1993). Although the portfolio's Sharpe ratio is reduced to 0.66 when trading cost is considered, the Sharpe ratio is still higher than the S&P 500 index between 2013 and 2022.

6.4 Embedding model

In this paper, we introduced a unique method to generate document embedding using two different NLP models; Doc2Vec and FinBERT.

We finalised on using both fine-tuned FinBERT and Doc2Vec models to create document embedding as it delivered the highest Sharpe ratio for the Hidden Neighbours portfolio. However, while the incorporation of FinBERT embedding may be suitable for our task of finding nuanced industry relationships, it does not imply that FinBERT or other transformer-based models must be used in constructing industry networks.

Rather, while FinBERT has shown superior performance in various finance NLP tasks (Huang et al. 2022), it performs worse than Doc2Vec embeddings when used alone to build the Hidden Neighbours portfolio, as shown in Table 6. The relatively poor performance could be due to the limits of compression and the low signal-to-noise ratio when applying transformer-based language models on long 10-X documents. Conversely, while simple document embedding methods better reproduce standard ICS relationships than machine learning-based models (He et al. 2020), it is unable to fully capture nuanced similarities between companies, which could be better captured with transformer-based models.

Overall, we believe that there is no one-size-fits-all document embedding model to depict industry relationships between companies. In the context of the Hidden Neighbours portfolio, using both FinBERT and Doc2Vec models delivered superior performance, showcasing how complex language models can be used in conjunction with bag-of-words embedding. However, we believe that the most appropriate embedding methodology depends on the scope and types of industry relationships the researcher wants to capture, requiring specialised experimentation with various NLP tools.

7 Conclusion

In summary, our study introduces a new methodology that merges complex network analysis and advanced NLP techniques for constructing industry momentum portfolios. The Combined Network framework leverages on incorporating differing information sets embedded in price-based and text-based networks, revealing less visible

relationships between S&P 500 companies. The resulting industry momentum portfolio demonstrates exceptional performance, boasting a Sharpe ratio surpassing the S&P 500 index, and other momentum benchmarks constructed solely on past returns or SIC industry classification.

Furthermore, we contribute a novel approach to textual information integration by combining FinBERT with Doc2Vec to generate document-level embeddings from corporate disclosures of S&P 500 companies. Based on the Combined Network, our results on the Hidden Neighbours portfolio deliver a higher Sharpe ratio of 0.85 than other benchmarks with a statistically significant Cahart's α and a moderate maximum drawdown.

The outcomes of our research hold significant implications for portfolio management and financial decision-making, offering a holistic strategy that embraces diverse data dimensions. We acknowledge the exploratory nature of this study, inviting further investigations into the ways of building stock networks potentially using alternative data sources, large language models, and portfolio optimisation techniques.

Our work underscores the potential of interdisciplinary methodologies, uniting financial theory, network science, and Natural Language Processing. By seamlessly blending the Combined Network framework with text similarity and stock price correlation, we provide practitioners with an innovative toolkit to navigate contemporary financial complexities, thereby advancing both theoretical and practical aspects of portfolio management.

Appendix A Details on document embedding NLP models

Doc2Vec

We remove common stop words used in corporate disclosure by removing the top 100 most commonly used words in Loughran-McDonald Master Dictionary (Loughran and McDonald 2011). Afterwards, we follow the implementation used in Adosoglou et al. (2022) to train our Doc2Vec model, which is a PV-DM implementation with 256-dimensional embedding, and is trained for 10 epochs. If one master Doc2Vec model is trained for all years, there is room for look-ahead bias. Thus, for each calendar year, we separately trained a new Doc2Vec model using the calendar year's 10-K and 10-Qs. Afterwards, the calendar year's Doc2Vec model was used to create document embedding for 10-K and 10-Qs. Finally, an annual vector representation was created by averaging the four document embeddings, one 10-K and three 10-Qs.

FinBERT

We generate document embedding on 10-X disclosures using a Sentence BERT implementation of pre-trained FinBERT Huang et al. (2022) with mean pooling.

This model is fine-tuned using SimCSE (Gao et al. 2021) method. SimCSE is a contrastive learning process that does not need any labelling process.

Figure 1 shows the fine-tuning process of our transformer model using SimCSE. We first separated the 10-X document into constituent sentences. Afterwards, we duplicate each sentence once to generate a pair of identical sentences. Then, token embedding is generated for each word in a sentence using FinBERT. We then average the token embedding using Sentence BERT (Reimers and Gurevych 2019) mean-pooling operation to generate a sentence embedding with a dimension of 784.

Due to the dropout layer (Devlin et al. 2019; Huang et al. 2022) in FinBERT's transformer architecture, even when two identical sentences are inputted, the resulting sentence embeddings are slightly different to each other (Gao et al. 2021), indicated by u and u' in Fig. 1. This difference is relatively smaller when compared to a vector representation generated from a truly different sentence v from a different document. The SimCSE learning methodology fine-tunes the FinBERT model by maximising the cosine distance between truly different sets of sentences, u and v, while minimising the cosine distance between embeddings generated from the same sentences, u and u'.

Once completing the fine-tuning, it is necessary to use the FinBERT model to yield the vector representation for each 10-X document. To do so, the FinBERT model generates sentence embeddings for sentences in the respective 10-X document. Then, the sentence embeddings are averaged to generate our final document embedding with a dimension of 784. Similar to Doc2Vec method, we average the 10-X disclosures to create an annual representation.

Appendix B Disparity filter backboning method

The disparity filter backboning method (Serrano et al. 2009) effectively performs a statistical test on each edge weight assuming a null model, where the edge weights are assumed to be uniformly randomly distributed.[5] Serrano et al. (2009) recommends using significance level cut-off, such as 5% or 10%, commonly used in various statistical tests. This would translate into 0.95 and 0.90 cut-off for our $1-\alpha$ statistics.

However, the majority of our network edges are below 0.95 cut-off. This is because the disparity filter was originally intended for real-life social networks where each edge will represent a material connection between each nodes. However, in our application of disparity filter, our networks begin in a fully connected state where all nodes are connected to each other. Thus, in our calculation of normalised edge weight p_{ij}:

$$p_{ij} = \frac{w_{ij}}{\Sigma_i w_i},$$

[5] A Python implementation of disparity filter backboning can be found in Yassin et al. (2023).

where the denominator, $\Sigma_i w_i$, becomes inflated as every edge weight gets taken into account. This greatly reduces our $1-\alpha$ statistics value, to a threshold lower than 0.95 or 0.90.

While the application of original cut-off levels suggested by Serrano et al. (2009) is difficult, the disparity filter can still be used to rank the relative statistical significance of each edge weight. Thus, we continued to use this backboning methodology, with an intent to compare the relative statistical significance of each edge weight compared to the null model.

We adjust the cut-off point for $1-\alpha$ test statistics such that it is as large as possible while not leading to a loss of nodes. While it is ideal to set $1-\alpha$ as high as 0.95, only retaining statistically significant edges, this leads to loss of nodes which is not ideal for portfolio optimisation. We found 0.7 as the ideal cut-off level where statistically significant edges remain while no nodes are lost.

Appendix C Details and consideration around benchmark portfolio construction

For our Standard Momentum benchmark, we chose the 6-month holding period with 6-1 look-back period to ensure that the specification is the same as the SIC Industry Momentum benchmark.

The construction of our SIC Industry Momentum differs from the original construction methodology suggested by Moskowitz and Grinblatt (1999). The main difference is that among the SIC Industry constituents experiencing momentum, we rank the top 50 stocks with highest total returns. This additional step was necessary as we had to ensure that the benchmark has the same number of stocks compared to the Hidden Neighbours portfolio. Due to the additional filtering, our SIC Industry Momentum benchmark could be more concentrated compared to the Moskowitz and Grinblatt (1999) portfolio, but we do not believe this additional sorting of momentum led to a material decrease in performance.

Acknowledgements I thank Dr. Seonho Park and Dragos Gorduza for their guidance, support, and comments on this paper. I am grateful to the anonymous referees for their insightful comments and constructive suggestions, which significantly improved the quality of this manuscript. I also extend my thanks to the editor for their support and guidance throughout the revision process.

Funding Partial financial support was received from Oxford Internet Institute for providing access to computational resources, mainly GPU usage.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

References

Adosoglou, G., Lombardo, G., Pardalos, P.M.: Neural network embeddings on corporate annual filings for portfolio selection. Expert Syst. Appl. **164**, 114053 (2021)

Adosoglou, G., Park, S., Lombardo, G., Cagnoni, S., Pardalos, P.M., et al.: Lazy network: a word embedding-based temporal financial network to avoid economic shocks in asset pricing models. Complexity **2022**, 9430919 (2022)

Baltas, A.-N., Kosowski, R.: Improving time-series momentum strategies: The role of trading signals and volatility estimators. SSRN eLibrary (2012)

Barroso, P., Santa-Clara, P.: Momentum has its moments. J. Financ. Econ. **116**(1), 111–120 (2015)

Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the international AAAI conference on web and social media **3**, 361–362 (2009)

Behr, P., Guettler, A., Truebenbach, F.: Using industry momentum to improve portfolio performance. J. Bank. Financ. **36**(5), 1414–1423 (2012)

Birch, J., Pantelous, A.A., Soramäki, K.: Analysis of correlation based networks representing DAX 30 stock price returns. Comput. Econ. **47**, 501–525 (2016)

Carhart, M.M.: On persistence in mutual fund performance. J. Financ. **52**(1), 57–82 (1997)

Chui, A.C., Titman, S., Wei, K.J.: Individualism and momentum around the world. J. Financ. **65**(1), 361–392 (2010)

Cohen, L., Malloy, C., Nguyen, Q.: Lazy prices. J. Financ. **75**(3), 1371–1415 (2020)

Daniel, K., Hirshleifer, D., Subrahmanyam, A.: Investor psychology and security market under-and overreactions. J. Financ. **53**(6), 1839–1885 (1998)

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)

Dyer, T., Lang, M., Stice-Lawrence, L.: The evolution of 10-K textual disclosure: evidence from latent Dirichlet allocation. J. Account. Econ. **64**(2–3), 221–245 (2017)

Faizliev, A., Balash, V., Petrov, V., Grigoriev, A., Melnichuk, D., Sidorov, S.: Stability analysis of company co-mention network and market graph over time using graph similarity measures. J. Open Innovation Technol. Mark. Complex. **5**(3), 55 (2019)

Fama, E.F.: Efficient capital markets: a review of theory and empirical work. J. Financ. **25**(2), 383–417 (1970)

Frazzini, A., Israel, R., Moskowitz, T. J.: Trading costs. Available at SSRN 3229719 (2018)

Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 (2021)

Gleason, K.C., Mathur, I., Peterson, M.A.: Analysis of intraday herding behavior among the sector ETFs. J. Empir. Financ. **11**(5), 681–694 (2004)

Grobys, K., Kolari, J.: On industry momentum strategies. J. Financ. Res. **43**(1), 95–119 (2020)

Grundy, B.D., Martin, J.S.M.: Understanding the nature of the risks and the source of the rewards to momentum investing. Rev. Financ. Stud. **14**(1), 29–78 (2001)

He, J., Chen, K., et al.: Exploring machine learning techniques for text-based industry classification (2020)

Hoberg, G., Phillips, G.: Text-based network industries and endogenous product differentiation. J. Political Econ. **124**(5), 1423–1465 (2016)

Hoberg, G., Phillips, G.M.: Text-based industry momentum. J. Financ. Quant. Anal. **53**(6), 2355–2388 (2018)
Hong, H., Lim, T., Stein, J.C.: Bad news travels slowly: size, analyst coverage, and the profitability of momentum strategies. J. Financ. **55**(1), 265–295 (2000)
Huang, A.H., Wang, H., Yang, Y.: FinBERT: a large language model for extracting information from financial text. Contem. Account. Res. **40**(2), 806–841 (2022)
Jegadeesh, N., Titman, S.: Returns to buying winners and selling losers: implications for stock market efficiency. J. Financ. **48**(1), 65–91 (1993)
Jeon, S. W., Lee, H. J., Cho, S.: Building industry network based on business text: corporate disclosures and news. In: 2017 IEEE International Conference on Big Data (Big Data), pages 4696–4704. IEEE (2017)
Kimura, Y., Nakagawa, K.: Industry momentum strategy based on text mining in the japanese stock market. In: 2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI), pages 420–423. IEEE (2022)
Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning, pages 1188–1196. PMLR (2014)
Lee, J.W., Nobi, A.: State and network structures of stock markets around the global financial crisis. Comput. Econ. **51**, 195–210 (2018)
Lewellen, J., Nagel, S., Shanken, J.: A skeptical appraisal of asset pricing tests. J. Financ. Econ. **96**(2), 175–194 (2010)
Li, S.: Industry classification, industry momentum and short-term reversal. Financ. Res. Lett. **48**, 102860 (2022)
Li, X., Brooks, C., Miffre, J.: Low-cost momentum strategies. J. Asset Manag **9**, 366–379 (2009)
Loughran, T., McDonald, B.: When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. J. Financ. **66**(1), 35–65 (2011)
Loughran, T., McDonald, B.: Textual analysis in finance. Annual Rev. Financ. Econ. **12**, 357–375 (2020)
Loukas, L., Fergadiotis, M., Androutsopoulos, I., Malakasiotis, P.: Edgar-corpus: Billions of tokens make the world go round. arXiv preprint arXiv:2109.14394 (2021)
Mantegna, R.N.: Hierarchical structure in financial markets. European Phys. J. B-Condens. Matter Complex Syst. **11**, 193–197 (1999)
Marti, G., Nielsen, F., Bińkowski, M., Donnat, P.: A review of two decades of correlations, hierarchies, networks and clustering in financial markets. Progress Inf. Geom. Theory Appl., pages 245–274 (2021)
Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
Moskowitz, T.J., Grinblatt, M.: Do industries explain momentum? J. Financ. **54**(4), 1249–1290 (1999)
O'Neal, E.S.: Industry momentum and sector mutual funds. Financ. Anal. J. **56**(4), 37–49 (2000)
Phillips, R.L., Ormsby, R.: Industry classification schemes: an analysis and review. J. Bus. Financ. Librariansh. **21**(1), 1–25 (2016)
Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
Serrano, M.Á., Boguná, M., Vespignani, A.: Extracting the multiscale backbone of complex weighted networks. Proc. Natl. Acad. Sci. **106**(16), 6483–6488 (2009)
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
Wiest, T.: Momentum: what do we know 30 years after Jegadeesh and Titmanâ€™s seminal paper? Financ. Mark. Portf. Manag. **37**(1), 95–114 (2023)
Yassin, A., Haidar, A., Cherifi, H., Seba, H., Togni, O.: Netbone: A python package for extracting backbones of weighted networks. In: French Regional Conference on Complex Systems (2023)

Joon Chul James Ahn holds an MSc from the University of Oxford. His research interests include the use of machine learning tools to solve complex finance problems, as well as other topics in quantitative social science. Currently, he is working as an economic consultant.

Dragos Gorduza is a Ph.D. student at the University of Oxford's Oxford-Man Institute of Quantitative Finance. His research interests include the use of Graph Neural Networks and Natural Language Processing tools to solve complex finance problems. He has worked as a research staff member at various institutions, including the Bank of England and the Alan Turing Institute.

Seonho Park is a Postdoctoral Fellow at the Georgia Institute of Technology. Prior to that, he holds a Ph.D. degree in industrial and systems engineering from the University of Florida. His research primarily revolves around optimisation and machine learning in applications of power system and finance. Currently, he is specifically focused on utilising machine learning techniques to accelerate the optimisation process in power system applications and LLM techniques for quantitative research in finance.