Hindawi
Complexity
Volume 2022, Article ID 9430919, 12 pages
https://doi.org/10.1155/2022/9430919

WILEY | Hindawi

Research Article

Lazy Network: A Word Embedding-Based Temporal Financial Network to Avoid Economic Shocks in Asset Pricing Models

George Adosoglou,[1] **Seonho Park**,[1] **Gianfranco Lombardo**,[2] **Stefano Cagnoni**,[2] **and Panos M. Pardalos**[1,3]

[1]*Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA*
[2]*Department of Engineering and Architecture, University of Parma, Parma, Italy*
[3]*LATNA, Higher School of Economics, Moscow, Russia*

Correspondence should be addressed to Gianfranco Lombardo; gianfranco.lombardo@unipr.it

Received 6 December 2021; Accepted 11 March 2022; Published 22 April 2022

Academic Editor: Guilherme Ferraz de Arruda

Public companies in the US stock market must annually report their activities and financial performances to the SEC by filing the so-called 10-K form. Recent studies have demonstrated that changes in the textual content of the corporate annual filing (10-K) can convey strong signals of companies' future returns. In this study, we combine natural language processing techniques and network science to introduce a novel 10-K-based network, named Lazy Network, that leverages year-on-year changes in companies' 10-Ks detected using a neural network embedding model. The Lazy Network aims to capture textual changes derived from financial or economic changes on the equity market. Leveraging the Lazy Network, we present a novel investment strategy that attempts to select the least disrupted and stable companies by capturing the peripheries of the Lazy Network. We show that this strategy earns statistically significant risk-adjusted excess returns. Specifically, the proposed portfolios yield up to 95 basis points in monthly five-factor alphas (over 12% annually), outperforming similar strategies in the literature.

1. Introduction

A 10-K form is an annual report that is required by the Securities and Exchange Commission (SEC) for publicly traded companies and becomes a reliable information source for many investors. The in-depth overview of a firm's annual performance mandated by the SEC in a 10-K renders it the most comprehensive corporate disclosure that has a regular cadence and continuous usage. However, due to an enormous amount of 10-Ks filed annually as well as ever increasing usage of redundant and boilerplate words, it has become exceedingly difficult for investors to scrutinize [1].

Cohen et al. demonstrated in their paper about "Lazy Prices" that year-over-year changes in 10-Ks tend to be unattended for long periods of time and demonstrated that the companies that change their 10-Ks significantly tend to underperform whereas the companies that do not tend to change their 10-Ks outperform the market [2].

Specifically, in the Lazy Prices paper, the authors used the Bag-of-Words (BoW) model to track the year-on-year changes in the 10-Ks. However, the BoW model suffers from the curse of dimensionality. To overcome this issue, neural network embedding techniques such as Doc2Vec [3] have been used to represent the 10-K documents as vectors, such that eventually the cosine similarity between the representing vectors captures the year-on-year change in 10-Ks efficiently [4].

In this study, we extend this idea by measuring how much a firm's 10-K changes relative to its closely related companies' 10-Ks. This is motivated by the fact that yearly changes of the financial status of a company in the stock market can be captured not only by its own changes over years but also by its changes relative to other companies having similar characteristics. This is possible because of the unique nature of the universe of the 10-Ks each year where similar companies can be identified using the cosine

similarity between the document vector representations [5, 6] or just by clustering the embeddings [7]. Under the studied scenario, a company's 10-K changing relative to another company's 10-K or even a group of other 10-Ks can be measured as the year-on-year change in the total cosine similarity, i.e., the sum of the cosine similarities between the 10-K under consideration and all other 10-Ks.

In order to measure the yearly and relative financial changes of the companies to their rivalry companies, we propose the Lazy Network that is constructed in the following way: (1) We construct networks based on the 10-Ks and document embedding technique for each year where the edges represent the similarity between 10-Ks of the companies. (2) Then, the yearly difference of two consecutive networks can be calculated by the absolute differences between edge values in two consecutive years. (3) Finally, we construct the resultant network, termed Lazy Network, over two consecutive years; its edges correspond to the absolute difference values, and nodes represent the companies in the stock market.

After constructing the Lazy Network, it is used to identify the *peripheral area* in the network where there is not a lot of changes happening, compared to other companies. The companies in this area tend to have extraordinary stability against other companies in the same industry. Thus, we capture those companies and construct long-only portfolio strategies consisting of those that lead to statistically meaningful performance.

The name of Lazy Network is also inspired by Lazy Prices [2] in which the authors demonstrated that companies tend to be *lazy*; i.e., they just append some words to their last year's 10-K to create a current year's 10-K. They also tend to change only what is necessary to fulfill their fiduciary duties. The Lazy Network has been named after having the same property as Lazy Prices. On top of that, we also extend it to have (1) neural network-based text embedding and (2) network representation of the companies to construct the more sophisticated portfolios.

Overall, the main contributions of this paper are as follows:

(i) We show how to construct the Lazy Network that is basically based on (1) form 10-K, (2) neural network-based text embedding, and (3) network representation of the connectivity between the companies.

(ii) By exploiting the Lazy Network, we propose four portfolio strategies that give high future risk-adjusted excess returns.

(iii) We show that these portfolio strategies are superior to previous baselines by presenting their competitive performance over the better.

Section 2 deals with the works related to our approach, the Lazy Network. In Section 3, we revisit some methodological backgrounds including the natural language processing model we used to extract the vectors from the 10-Ks and the centrality measures we use to capture the relatively stable companies to be included in the portfolios. We present the details on the way of constructing the Lazy Network and its corresponding portfolio strategies in Section 4, followed by the performance results and discussions about those in Section 5. Lastly, we give some concluding remarks in Section 6.

2. Related Work

Numerous studies have used various types of data to document financial linkages and build financial networks [8]. The most common approach has been to use stock price correlations to determine network edges. This idea is originated by [9] that established the standard methodology used in many studies in the following two decades. Hoberg and Phillips [5] demonstrated that the stock price similarity changes can reflect positive and negative demand shocks occurring within an industry. For example, they showed that the positive demand shock in military industry because of the September 11 attacks can be captured by the increase in the similarity between the military companies because these companies attempted to fill the product demand gap. Moreover, a negative demand shock because of the post-2000 software market collapse decreased the similarity in the software industry as software companies attempted to pivot into other product markets.

Other types of financial networks have been also constructed. Based on the supply chain data, dynamic supplier networks were built, and the centrality measures for each company were computed [10]. The author found that stock price movements of central suppliers can forecast the following overall market performance. Cooccurrence networks have been created by [11], from parsed news articles, and used as risk indicators. Similarly, Chen et al. [12] constructed a media network based on investors' attention predicting abnormal returns. Souza and Aste [13] combined stock price data with past social media sentiment information in a network to predict the stock market structure.

Using director-based social networks, Houston et al. [14] analyzed soft and hard information and showed how centrality measures can aid in predicting partnerships and roles in syndicated lending. Most similar to this work, Jeon et al. [6] employed Doc2Vec on 10-Ks as well as Bloomberg news data to build industry networks.

Existing literature has focused on portfolio selection by utilizing financial network structures. Specifically, Pozzi et al. [15] computed hybrid centrality measures based on five common centrality measures and found peripheral securities having better returns with lower risk compared to central ones. Onnela et al. [16] also found that, during market crashes, the Minimum Spanning Tree (MST) [9] greatly contracts and that almost always the optimal Markowitz portfolios lie on the peripheries of the MST. They also found that there is a connection between peripheral stocks and optimal Markowitz portfolios and that it is preferable to include peripheral securities in a portfolio [17, 18]. The importance of the centrality of stocks has also been validated by Leone et al. [19] who showed that connected but low-degree centrality stocks outperform ones with high-degree centrality. Lee et al. [20] showed how the incorporation of

financial network indicators can enhance investment strategies by forecasting the global stock markets.

Other approaches to portfolio selection that do not use stock price correlation based networks have also been investigated. By constructing a network through trade relations, it was also found that the more central an industry is, the greater its exposure to sectoral shocks which transmit from one sector to another through trade [21]. By employing a PageRank-type algorithm [22], a dynamic measure of firm competitiveness has also been defined [23]. In particular, by constructing a network from firms' competitors mentioned in their 10-K filings, they found that a firm's competitiveness is coming primarily from its ability to compete across different business environments.

Also the centrality metrics are quite important to select firms for portfolio from the networks. Federico et al. [24] created a metric to measure the importance of nodes in the evolution of networks. Federico introduced the "change centrality" metric which motivated several of the metrics used in this study. This metric is a centrality measure in the sense that it quantifies how important a node is with respect to the network's changes.

3. Background

3.1. Neural Language Model. In natural language processing (NLP), neural network-based methods have been employed to extract salient information from various types of documents and texts for various tasks including translation and prediction. Neural network-based approaches mainly aim to establish a statistical model capable of predicting the joint probability of the words in the documents. A first attempt to do so was made by [25]. That paper proposed the first neural network-based language model, which is referred to as a probabilistic feedforward neural network language model or *Distributed Representation* and tries to predict a joint distribution of the words as well as represent each word as a vector in a lower-dimension space.

Recently, another prevalent method, namely, Word2Vec, has been introduced in [26]. This word embedding algorithm comprises a shallow neural network that has a linear activation function which maps words to distributed low-dimensional feature vectors. The Word2Vec model learns the feature vector of a word as the indirect result of predicting the context surrounding the other words (Continuous BoW) or the opposite trying to predict the missing word given the context (Skip-gram model). To achieve this result, the Skip-gram model of Word2Vec learns a probability distribution over the words and tries to minimize the average negative log of the probability by adjusting the neural network weights. The output layer of the neural network is a Hierarchical Softmax function. The equation that describes the Skip-gram objective function is the following:

$$\frac{1}{T}\sum_{t=k}^{T-k}\log\, p\left(w_t|w_{t-k},\ldots,w_{t+k}\right), \tag{1}$$

where

$$p\left(w_t|w_{t-k},\ldots,w_{t+k}\right)=\frac{e^{y_{w_t}}}{\sum_i e^{y_i}}. \tag{2}$$

The objective function is optimized using Stochastic Gradient Descent (SGD) in the form of backpropagation on a single hidden-layer feedforward neural network. One-hot encoded words are fed into the network, while the hidden layer has no activation function. The embedding of a word corresponds to the value in the embedding matrix that contains the weights between the one-hot encoded input and the hidden layer. In order to extend this framework, the authors come up with Doc2Vec [3] where a special paragraph token is considered in each document to embed directly the entire content of the document. The paragraph token vectors are learned with other word vectors during training. There are two prevalent variants of Doc2Vec, distinguished by the way of training: Distributed Bag-of-Words version of Paragraph Vector (PV-DBoW) and Distributed Memory version of Paragraph Vector (PV-DM). These, PV-DBoW and PV-DM modes, are extensions of the CBoW and Skip-gram of Word2Vec, respectively. In our work, we used the PV-DM mode of Doc2Vec for our experiments because of its better ability to capture semantic changes [4]. Similar techniques have been developed to perform node embedding in networks for node classification and link prediction tasks for static graph, e.g., Node2Vec [27]; for graph clustering optimization [28]; for knowledge graphs [29]; for multiplex networks [30]; and for dynamic networks [31].

3.2. Network Centrality Measures. In this subsection, we provide all the formulations for computing the network centrality measures used in this paper. Please note that all these measures take into account the edges' weights in their computation.

The network representation of the semantic space that corresponds to our 10-Ks can be thought of as an undirected network $G=(V,E)$ where the node set V represents the companies and the edge set E comprises weighted edges w_{ij} where $i, j \in V$. Please note that if there is no edge between node i and j, then $w_{ij}=0$.

3.3. Strength Centrality. The strength s_i of node i in the network is computed by summing the weights of its adjacent edges, namely,

$$s_i=\sum_{j\in N(i)} w_{ij}, \forall i \in V, \tag{3}$$

where $N(i)$ is the set of neighbor nodes of node i.

3.4. Eigenvector Centrality. The eigenvector centrality of a node i corresponds to the ith element of the eigenvector $\mathbf{v}$ associated with the maximum eigenvalue, λ, of the adjacency matrix $A \in \mathbb{R}^{n\times n}$. The elements of the adjacency matrix are the weighted edges between nodes. The eigenvector centrality is defined as follows:

$$A\mathbf{v}=\lambda\mathbf{v}. \tag{4}$$

3.5. Closeness Centrality. The closeness centrality is defined as the inverse of *farness*. Farness is the sum of the distances

to all other nodes. The distance between nodes can be represented by the shortest path between nodes, which is usually calculated by Dijkstra's algorithm [32].

The closeness centrality of node i, c_i, is thus defined as the reciprocal of the average shortest-path distance to node i over all the $n-1$ reachable nodes:

$$c_i = \frac{n-1}{\sum_{j=1, j\neq i}^{n} d(i,j)}, \tag{5}$$

where $d(i,j)$ is the shortest-path distance between node i and j.

3.6. Betweenness Centrality. The betweenness centrality of node i is defined as the total number of shortest paths between nodes that pass through node i:

$$b_i = \sum_{s,t\in V} \frac{\sigma(s,t|i)}{\sigma(s,t)}, \tag{6}$$

where $\sigma(s,t)$ is the number of the shortest paths between node s and t, and $\sigma(s,t|i)$ is the number of the paths passing through node i. If $s=t$, then $\sigma(s,t)=1$, and if $i\in\{s,t\}$, then $\sigma(s,t|i)=0$.

4. Methodology

The proposed methodology is summarized below:

(i) Data collection and selection: For all years from 2000 to 2018, all available SEC 10-K filings and their equivalents were collected along with their market capitalization and monthly returns. To avoid biases in our dataset as well as any extreme returns from outliers, we filtered our collection based on the market evaluation and annual return.

(ii) Preprocessing and model training: All the 10-Ks were preprocessed and cleaned from any tables, URLs, HTML tags, and stop words. Stemming was also applied to derive the root of each word and decrease the size of the dictionary. The PV-DM Doc2Vec model was afterwards trained on the entire document collection to get the 10-K filings' embeddings.

(iii) Construction of the Lazy Network: By means of the cosine similarity between the embedding vectors, we construct the 10-K Network for each year and then, by considering the absolute value of the differences between cosine similarity values for two consecutive years' 10-K Networks, we derive the Lazy Network proposed in this work.

(iv) Portfolio selection and evaluation: We construct four value-weighted and four equally weighted portfolios by leveraging four different centrality measures, namely, strength, eigenvector, closeness and betweenness centrality, in order to measure the impact of economic shocks on the companies in terms of the shocks' direct effects, total influence, immediacy, and mediative effects, respectively.

4.1. Data Collection and Selection. From the Loughran-McDonald dataset (https://sraf.nd.edu/data/stage-one-10-x-parse-data/), we collected all the 10-Ks and 10-K-related SEC filings (10-K, 10-KT, 10-KSB, 10-K405) from 2000 to 2018. We, then, collected all the monthly stock price data for the same firms from the Center for Research in Security Prices (CRSP). Because the SEC identifies companies only through their CIK codes, we used the WRDS linking tables. With this data, we computed the market capitalization and the monthly returns (including dividends).

For each year, we filtered the companies based on two policies: (1) We excluded nano-caps, companies whose market capitalization is below 50 million dollars. Nano-caps correspond to a big part of the publicly traded stocks; however, they are prone to yield extreme returns (see the discussion in [33]), resulting in affecting the analysis in our work negatively. (2) We also excluded firms whose annual return value exceeded 10,000% during the years we consider to avoid any outliers.

After this filtering, we totally gathered 830,664 firm-month observations. The resulting dataset gives equally and value-weighted returns quite similar to those of the CRSP stock universe, confirming that no bias of any kind has been introduced. We discuss this further in the "Results and Discussion."

4.2. Text Preprocessing and Model Training. Then, we trained the PV-DM Doc2Vec model for various vector dimensions and hyperparameters on the total corpus consisting of all companies' 10-Ks. To combine the vectors, we used concatenation when we chose the number of epochs and the size of the vector, and the rest of the hyperparameters were chosen following the recommendations of [34]. We ended up using the PV-DM Doc2Vec model that outputs a 256-dimensional embedding vector.

4.3. Construction of the 10-K Network. We then constructed consecutive yearly networks where the nodes represent companies and the edges represent the similarity between companies' 10-Ks. Specifically, the weighted edge w_{ij} is determined as $\max(CS(i,j),\theta)$ where $CS(\cdot,\cdot)$ represents the cosine similarity between the embedding vectors generated by the Doc2Vec model of two companies' 10-Ks and θ is a predefined threshold called the slicing cutoff.

It is known that a proper degree distribution of the real-world big network follows the power law [35]. By the term "degree" here, we mean the number of edges connected to the node, that decreases with a higher slicing cutoff. To appropriately select the value of the slicing cutoff so that the resultant network's degree distribution follows this law, we performed a threshold analysis with various slicing cutoff values.

Figure 1 illustrates, in the log-log scale, the degree distributions of the network for the first (2000), middle (2009), and last (2018) years. The total degree distribution of all the years is also illustrated. For the degree distributions, we show the output when using 0.3 to 0.6 as the slicing cutoffs. In order to choose the appropriate

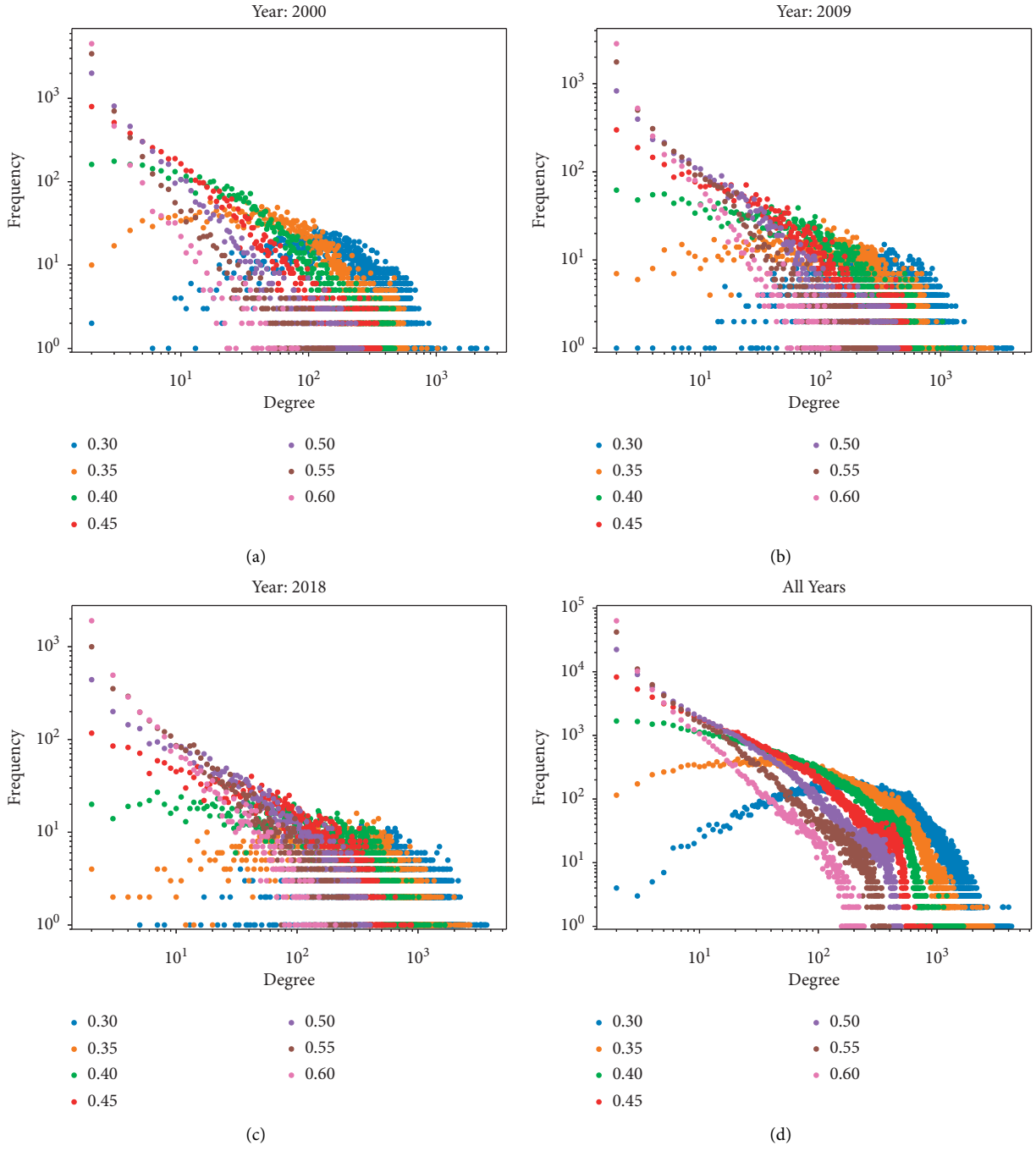

Figure 1: Node degree distributions for the first year (2000), middle year (2009), and last year (2018) 10-K Network as well as for all the nodes in all years from 2000 to 2018, sliced at seven different thresholds (0.3, 0.35, 0.4, 0.45, 0.5, 0.55, and 0.6).

slicing cutoff across the years, we investigated the Kolmogorov–Smirnov (KS) statistic and its associated p value while considering low-degree saturation [37]. Table 1 shows the mean and standard deviation values over the years. In this table, the KS statistic represents the maximum distance between the empirical degree distribution and its theoretical power law function. The smaller the KS statistic, the more closely the degree distribution follows the theoretical power law. In addition, as a *goodness-of-fit* test, the p value quantifies the plausibility of the hypothesis. According to [36], if the p value is larger than 0.1, the power law is a plausible hypothesis for the data at a significance level of 10%. Therefore, based on these experiments, we have chosen 0.4 as the slicing cutoff value for constructing the 10-K Network across all the years.

It is also important to mention that we have tested using network filtering techniques such as the Minimum Spanning Tree (MST) and Planar Maximally Filtered Graph (PMFG) [38] to construct the 10-K Networks because those filtering

Table 1: Average ± standard deviation of Kolmogorov–Smirnov (KS) statistic and its corresponding p value for different threshold values from 2000 to 2018. The higher the p value, the more plausible it is that the degree distribution follows the power law. According to [36], a p value larger than 0.1 provides strong evidence of following the power law at a significance level of 10%.

Threshold θ	0.30	0.35	0.40	0.45	0.50	0.55	0.60
KS statistic	0.17 ± 0.03	0.08 ± 0.04	0.05 ± 0.01	0.09 ± 0.04	0.10 ± 0.05	0.10 ± 0.03	0.14 ± 0.03
p value	0.00 ± 0.00	0.27 ± 0.32	0.62 ± 0.31	0.18 ± 0.21	0.19 ± 0.30	0.03 ± 0.06	0.01 ± 0.02

techniques are frequently exploited in financial network applications as mentioned in [15]. However, we could not find any statistically meaningful performances in terms of cumulative results. It is conjectured that using the filtering techniques can hinder the centrality measures from finding the periphery because it gets rid of some connectivity in the network.

4.4. Construction of the Lazy Network. Since companies update their 10-Ks every year, the 10-K Network is also evolving accordingly. Because these updates reflect the changes in the business situation of these companies, it is important to track these changes relative to each other, which can also help us study economic shocks in the market. In an effort to measure how much the companies are affected by various circumstances in the market, we construct the Lazy Network, which has weights corresponding to the absolute difference of each two consecutive years' adjacency matrices of the 10-K Networks. A visual representation of how the Lazy Network is constructed is given in Figure 2.

Specifically, in our case, larger weights of the Lazy Network represent 10-Ks becoming more dissimilar to each other. For example, a higher weight value of the Lazy Network means that the two companies linked by the corresponding node change quite differently over the years even though those two companies are similar in terms of the cosine similarity. Also note that if two companies are quite different from each other, the corresponding weight in the 10-K Networks is zero (meaning lower than the slicing cutoff), resulting in zero weight in the Lazy Network as well. Therefore, the center nodes of the Lazy Network represent companies that are relatively changing their 10-Ks quite a lot compared to their neighbor companies over two consecutive years. Based on the same rationale as Lazy Prices [2], when it comes to construct the portfolio, we expect that the companies in the center of the Lazy Network tend to underperform; thus, we capture the periphery of the Lazy Network and include those located in the periphery in the Lazy Network-based portfolios.

4.5. Portfolio Selection and Evaluation. After we construct the Lazy Network, we compute four centrality measures, namely, strength, eigenvector, closeness, and betweenness centrality, to capture the periphery of the Lazy Network. An illustrative example for the Lazy Network obtained with the different centrality measures is presented in Figure 3.

The reason for testing all four centrality measures separately and not using a hybrid measure is that it is actually common for a node to appear central in one centrality measure and peripheral in another as they reflect different criteria and capture different aspects of the topology of the network. This is also evident in Table 2 that shows the correlations across different centrality measures. The correlations in this table are in line with expectations, given these measures' definitions (see "Network Centrality Measures"). We observe that the betweenness centrality is the least correlated with the other three measures, which is expected since it measures a brokerage-like attribute of the nodes. On the other hand, the eigenvector centrality and closeness centrality are highly correlated, also as expected, since they both consider the full network. Finally, all measures are correlated with strength centrality to a certain degree, which is the simplest measure of centrality, merely capturing local connections.

After computing all types of centrality measures for all the years, we sort the companies based on them, and then for each year, for inclusion in the portfolio, we select the 50 peripheral companies with the lowest centrality measures. Specifically, we form eight long-only portfolios based on these four centrality measures and two basic portfolio weighting methods, equally and value-weighted. In the value-weighted portfolio, the 50 participants are weighted according to the beginning market capitalization, whereas in the equally weighted portfolio, the 50 participants are equally weighted. We then compute the corresponding calendar time portfolios and show their performance in terms of risk-adjusted excess returns as well as a comparison to other similar strategies in the literature in the following sections.

The holding period of all portfolios is 12 months (from April 1st to March 30th of the following year), with the rebalancing occurring every year on April 1st. The benchmarks used, the CRSP US Total Market Index and the S&P 500 Index, are rebalanced quarterly. We published in an external source the list of companies (CIK codes) selected for each centrality measure (http://sowide.unipr.it/sites/default/files/files/2022-lazy-network-portfolios.zip).

5. Results and Discussion

In this section, we present the performance of each of the four network centrality measures in capturing future abnormal returns. We also compare our performance to other two strategies that also use 10-Ks.

In these strategies, the 10-Ks are also represented as vectors in the first one ([2]) which uses the BoW model and the second ([4]) which uses the Doc2Vec model. After vectorizing the 10-Ks, these strategies go long on companies that features, the highest cosine similarity of their 10-K's vector representation to the previous year's

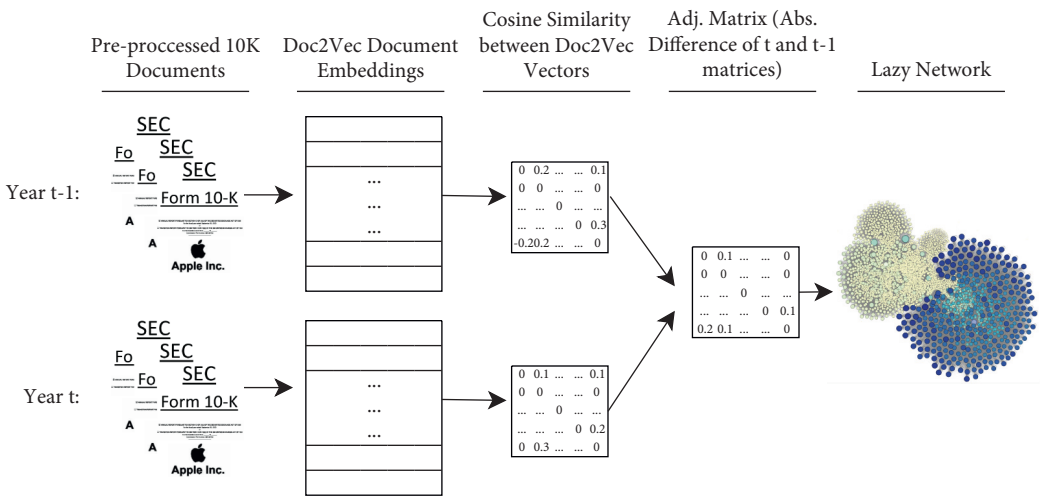

Figure 2: Flow chart for creating the Lazy Network for year t.

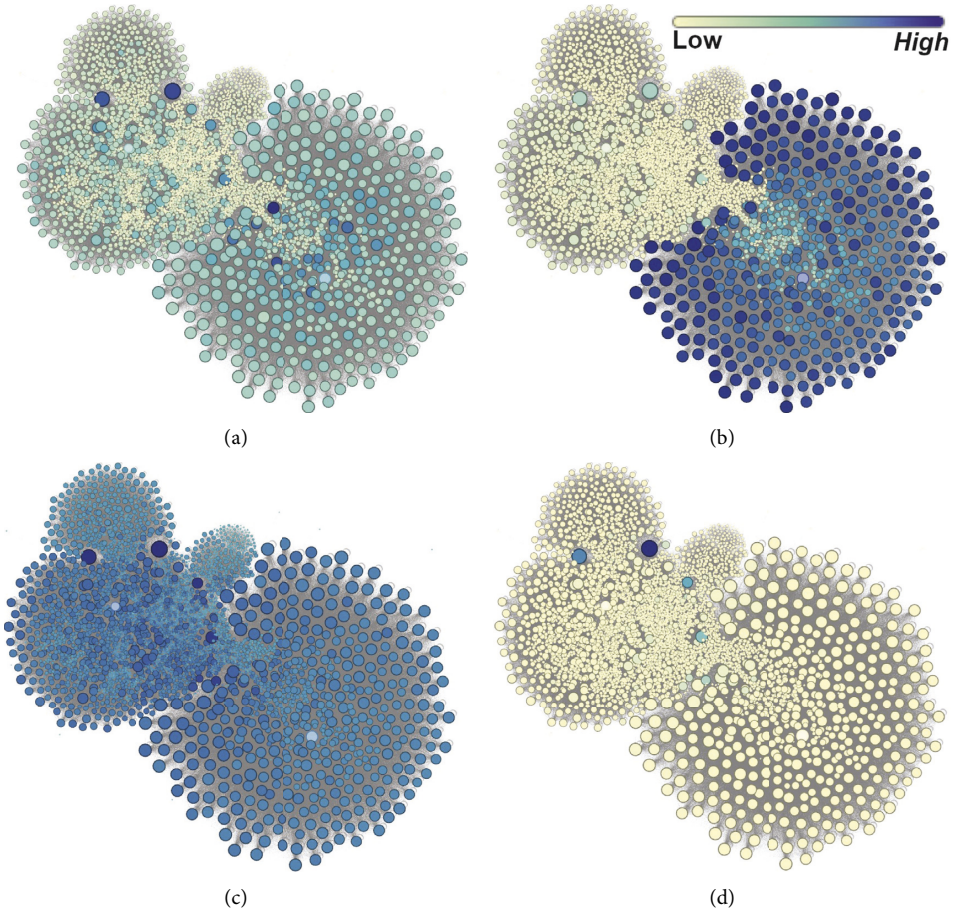

Figure 3: Lazy Network for the year 2014. In the network, the nodes are colored based on the different centrality measures while their size represents the degree. (a) Strength. (b) Eigenvector. (c) Closeness. (d) Betweenness.

Table 2: Mean ± standard deviation for the correlations of the network centrality measures for the years from 2000 to 2018.

Variable	Strength	Eigenvector	Closeness	Betweenness
Strength	1.000			
Eigenvector	0.81 ± 0.07	1.000		
Closeness	0.81 ± 0.04	0.88 ± 0.06	1.00	
Betweenness	0.59 ± 0.09	0.48 ± 0.09	0.65 ± 0.04	1.00

10-K. We believe it is appropriate to compare the three strategies because all three frameworks process the same input data and aim to capture market stability.

Finally, in Figure 4 we show the cumulative returns of all the portfolios constructed with each of the centrality measures from 2000 to 2019. We show two figures, one with the equally weighted and the other with the value-weighted cumulative returns. We compare these returns with buying and holding the S&P 500 Index, our available universe of stocks, and the CRSP stock universe. Note that dividends are included in all of the reported returns.

5.1. Asset Pricing Models. Because the large excess returns found in this study could have been a product of big systematic factor exposures, we also examine the returns with respect to the Fama–French 3-factor model [39] and 5-factor models [40, 41] that consider the Small Minus Big (SMB) size, the High Minus Low (HML) value factors as well as the Up Minus Down (UMD) momentum factor and Pástor–Stambaugh's traded liquidity factor (PS_VWF) in order to explain returns.

We empirically estimate the Capital Asset Pricing Model (CAPM) [33] and Fama–French 3-factor and 5-factor alphas and betas by linearly regressing the portfolios' monthly abnormal returns on these factors' monthly abnormal returns over the period from 2000 to 2018, as follows:

Capital Asset Pricing Model (CAPM):

$$R_t - r_t^f = \alpha + \beta_{\mathrm{MKT}}\left(R_t^M - r_t^f\right) + \varepsilon_t. \tag{7}$$

Fama–French 3-factor model:

$$R_t - r_t^f = \alpha + \beta_{\mathrm{MKT}}\left(R_t^M - r_t^f\right) + \beta_{\mathrm{HML}}\mathrm{HML}_t + \beta_{\mathrm{SMB}}\mathrm{SMB}_t + \varepsilon_t. \tag{8}$$

5-factor model:

$$R_t - r_t^f = \alpha + \beta_{\mathrm{MKT}}\left(R_t^M - r_t^f\right) + \beta_{\mathrm{HML}}\mathrm{HML}_t + \beta_{\mathrm{SMB}}\mathrm{SMB}_t + \beta_{\mathrm{UMD}}\mathrm{UMD}_t + \beta_{\mathrm{PS\_VWF}}\mathrm{PS\_VWF}_t + \varepsilon_t, \tag{9}$$

where $R_t - r_t^f$ is the excess return of a portfolio, $R_t^M - r_t^f$ is the market risk premium, and r_t^f is a risk-free rate of return. Here, we denote one-month Treasury rate as r_t^f. HML_t is constructed from the returns of the firms having the high book-to-market value minus the returns of the firms having low book-to-market value. SMB_t denotes the returns of small capitalization firms minus the returns of firms of large capitalization. UMD_t is the momentum factor, i.e., the difference between the returns of winners' and losers' portfolios considering the past 11 months. $\mathrm{PS\_VWF}_t$ is the Pástor-Stambaugh liquidity traded factor, which is the difference of the top liquidity beta decile portfolio returns minus the returns of the bottom decile portfolios—we do not subtract the risk-free rate from the SMB, HML, UMD, or PS_VFW portfolios, as these factors' returns are constructed from the differences between two portfolios, and thus the risk-free rate is already canceled out. The parameter α corresponds to the excess risk-adjusted return representing the abnormal return over what is expected merely on the basis of the riskiness of the portfolio. All the time series for the factor returns and the risk-free rates are collected from Wharton Research Data Services (WRDS; https://wrds-web.wharton.upenn.edu/wrds/) and the Fama–French datasets of portfolios and factors (https://mba.-tuck.dartmouth.edu/pages/faculty/ken.french/data_-library.html; the momentum factor (UMD) is referred to as MOM in the data library).

5.2. Performance Evaluation. The performance of each of the eight long-only portfolios that were constructed based on the four centrality measures (four equally and four value-weighted) are presented in Table 3. This table shows the monthly abnormal portfolio returns as well as their statistical significance. To compute these, we have regressed for each case the excess monthly returns on the market, the SMB, HML, UMD, and PS_VWF factors from 2000 to 2018. We select 50 companies for each year.

Table 3 shows very positive and statistically significant excess returns, 3-factor, and 5-factor alphas for almost all the constructed portfolios. We also observe that the equally weighted portfolio returns are similar in magnitude, but larger than the value-weighted results. This spread is the largest in the eigenvector case, while the only exception to this observation is the portfolio constructed using the closeness centrality where the value-weighted returns are larger than the equally weighted ones. Furthermore, the best performance among all the portfolios is that of the eigenvector centrality when the portfolios are weighted equally and that of the closeness centrality when they are value-weighted. Specifically, the portfolio using eigenvector centrality earns a large and significant 5-factor alpha of 96 basis points per month ($t = 4.69$) in the equally weighted case, while the portfolio using closeness centrality earns a 5-factor alpha of 92 bps per month ($t = 4.22$) in the equally weighted case.

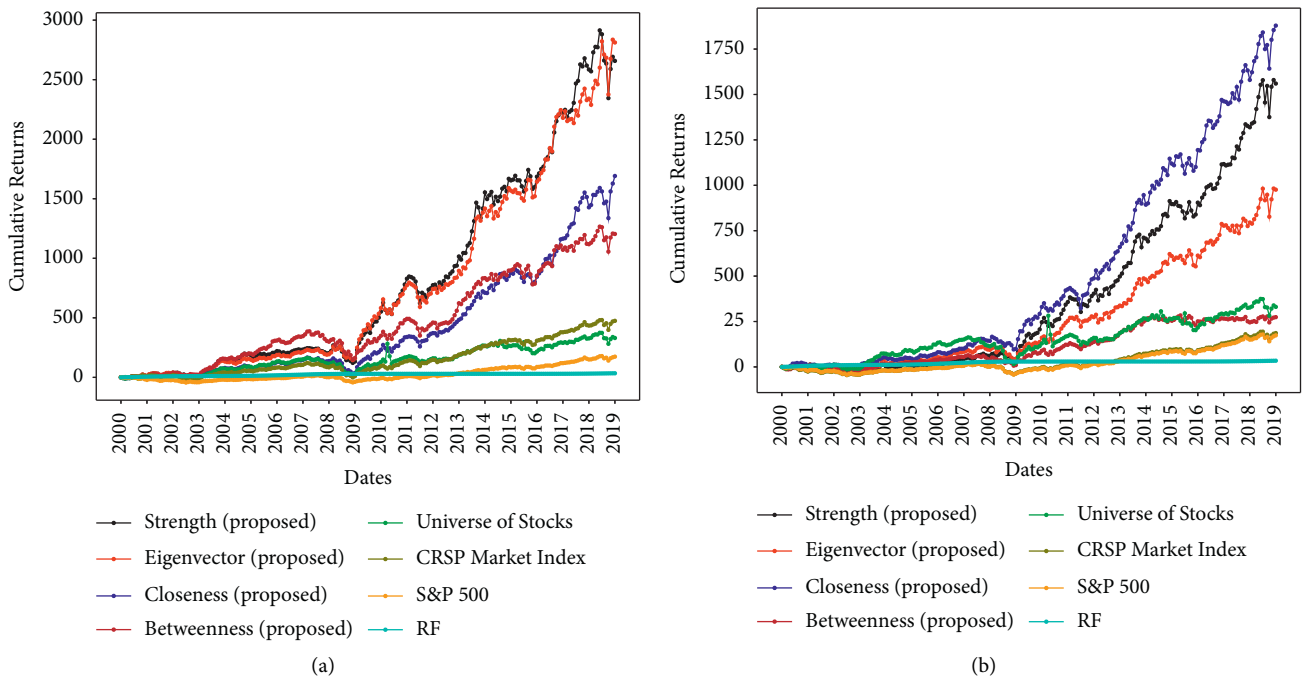

Figure 4: Equally weighted and value-weighted (left and right, respectively) cumulative returns from 2000 to 2019 for all the proposed portfolios using strength, eigenvector, closeness, or betweenness. The returns are compared to buying and holding the S&P 500 Index, as well as the whole universe of the available stocks portfolio (universe of stocks), and the CRSP market index. (a) Equally weighted cumulative returns. (b) Value-weighted cumulative returns.

Table 3: Monthly portfolio returns: monthly portfolio abnormal returns, 3-factor alphas, and 5-factor alphas (market, size, value, momentum, and liquidity) of four long-only portfolios that select the 50 most peripheral companies in terms of strength, eigenvector, closeness, and betweenness centrality in the Lazy Network.

Portfolio	Equally weighted			Value-weighted		
	CAPM alpha	3-factor alpha	5-factor alpha	CAPM alpha	3-factor alpha	5-factor alpha
Strength t-stat	1.05*** (4.21)	0.89*** (3.94)	0.94*** (4.38)	0.84*** (3.51)	0.76*** (3.23)	0.87*** (3.75)
Eigenvector t-stat	1.07*** (4.57)	0.92*** (4.35)	0.96*** (4.69)	0.64*** (2.62)	0.56** (2.30)	0.62** (2.52)
Closeness t-stat	0.84*** (4.32)	0.70*** (4.03)	0.72*** (3.80)	0.94*** (3.97)	0.90*** (4.14)	0.92*** (4.22)
Betweenness t-stat	0.71*** (4.08)	0.52*** (4.13)	0.51*** (4.07)	0.19 (1.08)	0.14 (0.83)	0.13 (0.78)

The t-statistics are in parentheses and displayed under the estimates. *, **, ***The significance at the 1%, 5%, and 10% levels, respectively.

We notice that the results are not affected when we control for the three Fama–French factors (market, size, and value) and the additional momentum and liquidity factors. This suggests that the returns observed are not explained by systematic exposures to common risk factors.

Only when using the betweenness centrality, the value-weighted portfolio failed to produce statistically significant alphas. However, this is acceptable, since betweenness centrality captures different attributes of the nodes that have to do mainly with the control of the flow between most other nodes. Furthermore, from Table 2 where the correlations between the centrality measures are illustrated, we see that the betweenness centrality shows the least correlation with the rest of the network centrality measures. On the other hand, the node strength centrality, even though the simplest centrality measure examined here, is still very illuminating, allowing the portfolio that uses it to yield alphas at the 0.90 level ($t = 4.38$). It is worth mentioning that in contrast to previous literature that examines, often using correlation networks, the association of centrality measures to the exposure to shocks or systemic risk contribution in general ([18, 21]), we do not detect the resiliency or susceptibility of companies; instead, we detect signals about the level of impact from previous economic shocks that goes unnoticed by investors for long periods of time.

5.3. Factor Loadings. For the portfolio that uses the closeness centrality, we report in Table 4 all the factor exposures obtained from the time-series regressions utilizing the CAPM, 3-factor, and 5-factor models. These loadings measure the exposures to the market, size, value, momentum, and liquidity risks. We observe statistically significant low betas in all portfolios ranging from 0.81 to 0.95. We also observe a big value tilt across the portfolios (the statistically

Table 4: Factor exposures of the portfolio that uses closeness centrality: the factor loadings of the long-only portfolio that goes long on the 50 companies with the lowest closeness centrality in the Lazy Network.

Factors	Equally weighted		Value-weighted	
	3-factor	5-factor	3-factor	5-factor
Intercept (α) t-stat	0.70*** (3.80)	0.72*** (3.97)	0.90*** (4.14)	0.92*** (4.22)
MKTRF t-stat	0.97*** (21.71)	0.90*** (18.92)	0.87*** (16.55)	0.81*** (14.30)
SMB t-stat	0.52*** (7.28)	0.52*** (7.40)	-0.03 (-0.35)	-0.03 (-0.30)
HML t-stat	0.26*** (4.38)	0.24*** (3.99)	0.17** (2.32)	0.15** (2.01)
UMD t-stat	—	-0.13*** (-3.41)	—	−0.11*
PS_VWF t-stat	—	0.07 (1.32)	—	0.06 (0.97)

The t-statistics are in parentheses and displayed under the estimates. *, **, ***The significance at the 1%, 5%, and 10% levels, respectively.

Table 5: Portfolio returns comparison: monthly excess returns (returns minus risk-free rate), 3-factor alphas (market, size, value), and 5-factor alphas (market, size, value, momentum, and liquidity) for the long-only portfolios, "Lazy Prices" [2] and "Semantic Similarity Portfolio," or "SSP" as referred to by [4]. To make comparisons easier, we also repeat the results from Table 3 for the proposed portfolios that use the eigenvector and closeness centrality. All these portfolios use similar rebalancing as well as similar datasets as input.

Portfolio	Equally weighted			Value-weighted		
	CAPM alpha	3-factor alpha	5-factor alpha	CAPM alpha	3-factor alpha	5-factor alpha
Eigenvector t-stat	1.07*** (4.57)	0.92*** (4.35)	0.96*** (4.69)	0.64*** (2.62)	0.56** (2.30)	0.62** (2.52)
Closeness t-stat	0.84*** (4.32)	0.70*** (4.03)	0.72*** (3.80)	0.94*** (3.97)	0.90*** (4.14)	0.92*** (4.22)
Lazy Prices [2] t-stat	0.96*** (3.05)	0.24*** (2.76)	0.23*** (2.70)	1.00*** (3.00)	0.44*** (2.78)	0.43*** (2.81)
SSP [4] t-stat	0.87** (2.56)	0.53** (2.50)	0.60** (2.69)	0.87** (2.75)	0.64** (2.27)	0.48 (1.59)

The t-statistics are in parentheses and displayed under the estimates. *, **, ***The significance at the 1%, 5%, and 10% levels, respectively.

significant HML loadings range from 0.15 to 0.49) as well as a statistically significant negative relation to the momentum factor (the UMD loadings range from -0.10 to -0.23).

These findings show that the portfolios yield lower risk, which is expected because these strategies have the defensive nature that avoids industry transformations. Defensive strategies usually tilt toward value. Our strategies avoid high beta, high growth, and high momentum stocks in rapidly changing industries. It is noted that text-derived information relates to the Fama and French value premium as well as the one-year momentum factor. The remaining factor loadings, SMB and PS_VWF, are generally not statistically significant. Note that the positive coefficient to the SMB factor in the equally weighted portfolios is expected given the weighting on smaller capitalization firms.

5.4. Performance Comparison. In this section, our results are compared to the results from other literature that uses companies' 10-Ks to predict future abnormal returns, namely, [2, 4]. In these studies, the changes are tracked only at the individual level, which means that networks are not being constructed. Specifically, these studies focus on companies that do not change their 10-K in significant way from the prior year. Before we proceed to the comparison, however, we document some differences between the data used for the portfolio selection in "Lazy Prices" and ours. (1) We examine the 2000–2018 period, while the authors of that paper study the period from 1995 to 2014. (2) The holding period is 12 months for our portfolios compared to 9 months for theirs. (3) They also consider companies that have "off-cycle" fiscal year-ends. (4) They select the quintile of firms that change their 10-Ks the least year over year in contrast to a constant number of companies selected in our strategy. We think that the comparison is fair because each portfolio is invested all year round, while the discrepancy in the periods should not be influencing the alphas generated by each strategy. No major differences in the data used exist when comparing to the strategy proposed in [4].

In Table 5, we show the performance of the portfolios found in literature that track only the year-on-year changes of the firms' 10-Ks. These are two long-only portfolios that go long on companies which do not change their 10-Ks in a significant way from year to year, one using the BoW model to represent the documents and the second using the PV-DM mode of Doc2Vec model. We denote the portfolios as "Lazy Prices" [2] portfolio and "Semantic Similarity Portfolio," or "SSP" as referred to by [4], respectively. In a direct comparison to the proposed portfolios constructed using the eigenvector and closeness centrality, we see that our portfolios outperform both strategies suggested in [2, 4] in terms of both equally and value-weighted portfolios' 3-factor and 5-factor alphas. Specifically, an equally weighted portfolio selecting companies that do not alter their 10-Ks in a significant way using the BoW model earns 23 basis points in monthly 5-factor alpha (a value-weighted portfolio earns 43 bps) and the one using the PV-DM mode of Doc2Vec earns 60 bps (a value-weighted portfolio earns 48 bps). All these 5-factor alphas are smaller both in value and statistical significance compared to the proposed portfolios. This outperformance is at the 40 bps level in monthly risk-adjusted excess returns.

5.5. Cumulative Returns. Figure 4 illustrates the cumulative returns of the equally and value-weighted portfolios based on the Lazy Network and the different centrality measures. It

also shows the benchmarks including the returns for the whole universe of stocks we consider, the CRSP market index, the S&P500 returns, and the risk-free rate. Firstly, we notice that there is no significant deviation of the available universe of stocks' returns from the CRSP market index returns, implying that the data used in this research is representative and it is free of survivor-biases or any other biases of any sort.

The widening gap between the performance of the suggested portfolios and the universe of stocks reflects the increasing complexity of the markets. Even though technological advancements have reduced the cost of information, relations between businesses are much harder to monitor and keep track of. Thus, investors are inattentive to this valuable information, which is then only reflected in prices after a substantial delay. As has been also evident in recent years, tech companies have the potential to negatively impact some of the non-tech companies. Companies that operate in areas with limited nascent competition might be outperforming areas with rising competition when we look at the market from a text-based network point of view.

Overall, the results in this paper highlight a distinct source of mispricing stemming from the slow reaction of investors to information about the effects of economic shocks and their propagation inside a network constructed with the similarities between firms' 10-K fillings. It is possible to determine the exact stocks that are most reflective of these market behaviors using the centrality measures in this network. It is important to stress that the non-disrupted companies are determined not only by their own 10-Ks but also by firms that are peers, suppliers, or customers by considering their adjacent nodes. This means that this is not just an independent evaluation based on attributes that are specific to a firm, such as size of the company or profitability, or even the competitive tone in its 10-K; instead, it captures the overall view among all companies concerning market events. This factor thus brings up the question as to whether investors are aware of the level at which the propagation of economic shocks is impinging on firms' performance.

6. Conclusion

In this study, we have developed a useful new metric of corporate risk from economic shocks that can be captured by analyzing the textual part of companies' 10-Ks. By leveraging neural network-based embedding techniques to construct the Lazy Network, we find that companies that are the least associated with economic shocks tend to outperform the market in a significant way. We argue that this happens because they face the fewest problems, as well as because these signals of stability are being overlooked by investors for long periods of time.

Specifically, the proposed novel portfolios that go long on companies which correspond to the least central nodes in the Lazy Network, measured by the strength, eigenvector, closeness, and betweenness centrality, yield relevant risk-adjusted excess returns. We find that the highest and most statistically significant 3-factor and 5-factor alphas are produced when we sort the companies based on the eigenvector and closeness centrality when constructing equally weighted portfolios and value-weighted portfolios, respectively. However, the portfolio that just sorts the stocks based on the strength centrality, which is the simplest measure in this study, also yields very high and statistically significant risk-adjusted excess returns of up to 90 basis points per month. Furthermore, we also found that even the worst portfolio based on betweenness centrality outperforms the baselines using 10-K.

The proposed approach can be used to detect emerging risks in specific sectors and industries and help identify the non-disrupted areas in the market where firms operate. The Lazy Network can be used for visual decision-making in financial management. Analysts, portfolio managers, retail investors, and policy makers alike could leverage the network to identify at any given moment which areas in the market are undergoing the greatest changes. Furthermore, our measures can be further applied to a variety of financial documents in which the timing is routine and the content is narrative and not restricted, for example, the proxy statements, the letters to shareholders, the earnings press releases, and the prepared part of earnings conference calls.

This study paves the way to research that integrates various approaches using deep learning and graph theory to analyze non-structured data, revealing valuable information to investors who are inattentive to them.

Data Availability

The data used for this research work are public and can be found in the SEC archive (https://www.sec.gov/cgi-bin/srch-edgar), for the less recent in the Loughran-McDonald dataset (https://sraf.nd.edu/textual-analysis/resources/).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work of P. M. Pardalos was conducted within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE). The work of G. Lombardo is partially financed by the Supercomputing Unified Platform Emilia-Romagna (SUPER) project, POR FSE, 2014/2020.

References

[1] T. Dyer, M. Lang, and L. Stice-Lawrence, "The evolution of 10-k textual disclosure: evidence from latent dirichlet allocation," *Journal of Accounting and Economics*, vol. 64, no. 2-3, pp. 221–245, 2017.

[2] L. Cohen, U. Axelson, and N. Barberis, "Lazy Prices 2019," *The Journal of Finance*, vol. 75, 2019.

[3] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," 2014, https://arxiv.org/abs/1405.4053.

[4] A. George, G. Lombardo, and P. M. Pardalos, "Neural network embeddings on corporate annual filings for portfolio

selection," *Expert Systems with Applications*, vol. 164, Article ID 114053, 2021.

[5] G. Hoberg and G. Phillips, "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, vol. 124, no. 5, pp. 1423–1465, 2016.

[6] W. J. Sung, H. J. Lee, and S. Cho, "Building industry network based on business text: corporate disclosures and news," in *Proceedings of the 2017 IEEE International Conference on Big Data, Big Data*, Boston, MA, USA, December, 2017.

[7] J. He and K. Chen, "Exploring machine learning techniques for text-based industry classification," 2020, https://ssrn.com/abstract=3640205.

[8] G. Marti, F. Nielsen, M. Bińkowski, and P. Donnat, "A review of two decades of correlations, hierarchies, networks and clustering in financial markets," *Signals and Communication Technology*, pp. 245–274, 2021.

[9] R. N. Mantegna, "Hierarchical structure in financial markets," *The European Physical Journal B*, vol. 11, no. 1, pp. 193–197, 1999.

[10] L. Wu, "Centrality of the supply chain network," 2015, https://ssrn.com/abstract=2651786.

[11] T. Forss and P. Sarlin, "News-sentiment networks as a risk indicator," 2017, https://arxiv.org/abs/1706.05812.

[12] Z. Chen, Li Guo, and J. Tu, "Media network and excess return comovement," *SSRN Electronic Journal*, 2019.

[13] T. T. P. Souza and T. Aste, "Predicting future stock market structure by combining social and financial network information," *Physica A: Statistical Mechanics and Its Applications*, vol. 535, Article ID 122343, 2019.

[14] J. F. Houston, J. Lee, and F. Suntheim, "Social networks in the global banking sector," *Journal of Accounting and Economics*, vol. 65, no. 2-3, pp. 237–269, 2018.

[15] F. Pozzi, T. Di Matteo, and T. Aste, "Spread of risk across financial markets: better to invest in the peripheries," *Scientific Reports*, vol. 3, no. 1–7, p. 1665, 2013.

[16] J. P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto, "Dynamics of market correlations: taxonomy and portfolio analysis," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 68, no. 5, Article ID 056110, 2003.

[17] Y. Li, X.-F. Jiang, Y. Tian, S.-P. Li, and B. Zheng, "Portfolio optimization based on network topology," *Physica A: Statistical Mechanics and Its Applications*, vol. 515, pp. 671–681, 2019.

[18] G. Peralta and A. Zareei, "A network approach to portfolio selection," *Journal of Empirical Finance*, vol. 38, pp. 157–180, 2016.

[19] A. Leone, M. Tomasini, Y. Al Rozz, and R. Menezes, "On the performance of network science metrics as long-term investment strategies in stock markets," *Studies in Computational Intelligence*, vol. 689, pp. 1053–1064, 2018.

[20] T. K. Lee, J. Hyung Cho, D. S. Kwon, and S. Y. Sohn, "Global stock market investment strategies based on financial network indicators using machine learning techniques," *Expert Systems with Applications*, vol. 117, pp. 228–242, 2019.

[21] R. Kenneth and Ahern, "Network centrality and the cross section of stock returns," *SSRN Electronic Journal*, 2013.

[22] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[23] A. Eisdorfer, K. Froot, G. Ozik, and R. Sadka, "Competition links and stock returns," *SSRN Electronic Journal*, 2019.

[24] P. Federico, J. Pfeffer, W. Aigner, S. Miksch, and L. Zenk, "Visual analysis of dynamic networks using change centrality," in *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, Istanbul, Turkey, August, 2012.

[25] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, https://arxiv.org/abs/1301.3781.

[27] A. Grover and J. Leskovec, "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, New York, NY, USA, August, 2016.

[28] L. Dong, Ru Yan, Q. Li, S. Wang, and J. Niu, "Semisupervised community preserving network embedding with pairwise constraints," *Complexity*, vol. 2020, Article ID 7953758, 14 pages, 2020.

[29] Yu. Zhao, J. Hou, Z. Yu, Y. Zhang, and Q. Li, "Confidence-aware embedding for knowledge graph entity typing," *Complexity*, vol. 2021, Article ID 3473849, 8 pages, 2021.

[30] R. Lu, P. Jiao, Y. Wang, H. Wu, and X. Chen, "Layer information similarity concerned network embedding," *Complexity*, vol. 2021, Article ID 2260488, 10 pages, 2021.

[31] G. Lombardo, A. Poggi, and M. Tomaiuolo, "Continual representation learning for node classification in power-law graphs," *Future Generation Computer Systems*, vol. 128, 2021.

[32] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[33] E. F. Fama and K. R. French, "Dissecting anomalies," *The Journal of Finance*, vol. 63, no. 4, pp. 1653–1678, 2008.

[34] H. L. Jey and B. Timothy, "An empirical evaluation of doc2vec with practical insights into document embedding generation," pp. 78–86, 2016, https://arxiv.org/abs/1607.05368.

[35] V. Boginski, S. Butenko, and P. M. Pardalos, "On structural properties of the market graph," *Innovations in financial and economic networks*, vol. 48, pp. 29–35, 2003.

[36] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[37] M. Morzy, T. Kajdanowicz, and P. Kazienko, "On measuring the complexity of networks: Kolmogorov complexity versus entropy," *Complexity*, vol. 2017, Article ID 3250301, 15 pages, 2017.

[38] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna, "A tool for filtering information in complex systems," *Proceedings of the National Academy of Sciences*, vol. 102, no. 30, Article ID 10421, 2005.

[39] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 1, pp. 3–56.

[40] M. M. Carhart, "On persistence in mutual fund performance," *The Journal of Finance*, vol. 52, no. 1, pp. 57–82, 1997.

[41] Ľ. Pástor and R. F. Stambaugh, "Liquidity risk and expected stock returns," *Journal of Political Economy*, vol. 111, no. 3, pp. 642–685, 2003.