Received: 16 January 2023 | Revised: 15 May 2023 | Accepted: 30 May 2023

DOI: 10.1002/isaf.1539

RESEARCH ARTICLE

WILEY

An efficient graph-based peer selection method for financial statements

Sander Noels[1,2] | Simon De Ridder[1] | Sébastien Viaene[1] | Tijl De Bie[2]

[1]Silverfin, Ghent, Belgium

[2]Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

Correspondence
Sander Noels, Silverfin, Ghent, Belgium.
Email: sander.noels@ugent.be

Funding information
Flanders Innovation & Entrepreneurship, Grant/Award Number: HBC.2020.2883; Flemish Government

Summary

Comparing companies can be useful for various purposes. Despite the widespread use of industry classification systems as a peer selection standard, these have been criticized for various reasons. Financial statements, however, offer a promising alternative to such classification systems. They are standardized, widely available, and offer deep insights into the nature of the company. In this paper, we present a graph distance metric for financial statements using the earth mover's distance. When using the distance metric on real-world tasks such as peer identification and industry classification, it shows promising results in terms of accuracy and computational efficiency.

KEYWORDS
company embedding, financial statements, graph distance metric, industry classification, peer companies

1 | INTRODUCTION

1.1 | Problem statement: company peer identification

Identifying peer companies is useful for a variety of reasons. It allows one to compare the performance of a company with its peers and identify areas of strength and weakness. Another key benefit is the ability to predict success. Companies that have comparable financial and operational performance to well-performing companies are likely to perform equally well (Hopkins, 1996). This allows for the identification of potentially successful companies. Additionally, identifying a company that deviates from its peers also contains interesting information. This may indicate a unique position with respect to its peers, which could be a sign of competitive advantage. On the other hand, it may also point to fraudulent behavior.

However, the process of peer identification can be challenging. Companies are often defined by a variety of financial and business characteristics, as well as their interactions with competitors, suppliers, customers, and joint ventures (Raman et al., 2019). This complexity makes it difficult to accurately identify peers. A thorough understanding of these factors is necessary for accurate peer identification.

1.2 | Industry classification systems

Currently, the task of selecting similar peers relies mostly upon industry classification systems, which are economic taxonomies that organize companies into economically related groups. However, this approach is not sufficient to accurately distinguish organizations. Companies provide a broad range of products and services and assigning them to a single category oversimplifies the company's business. Additionally, numerous companies are assigned to the same category, but the differences between those companies are too large to disregard. Therefore, there is a need for a more flexible system that uses a similarity score to quantify the resemblance between two companies, even if they have the same industry activity codes. By measuring the distance between two companies, it is possible to identify peer companies with greater granularity.

1.3 | Financial statements to characterize and compare companies

Financial statements provide a concise and comprehensive overview of a company's financial position and serve as a reliable predictor of future performance (Nagy & Obenberger, 1994). This encourages investors and government regulators to compare financial statements when making investment decisions or when detecting fraud. Furthermore, financial statements accurately depict business activities from a financial perspective. Although a large amount of financial statement data is available and offers a detailed summary of both the financial situation and the performed activities, it is not frequently utilized in the process of selecting peer companies. This is due to the reliance on labor-intensive manual processing and the lack of standardization.

In previous research, various attempts have been made to define company similarity based on financial statements for purposes such as company classification, fraud detection, and strategic peer selection (Brown et al., 2021; Hoitash et al., 2018; Jan, 2018; Kanapickienė & Grundienė, 2015; Yang & Cogill, 2013; Yang et al., 2019). However, these previously proposed methodologies only analyze a portion of the information present in a financial statement: either the values on the ledger accounts or the structural relationship between the ledger accounts present within a financial statement. This inspires the idea of considering both the structural properties and the value information to quantify the similarity between two companies. To the best of our knowledge, within the field of peer company selection based on financial statements, no methods consider both structure and value information.

1.4 | Motivation

In this paper, we present a novel approach to quantifying the similarity between two financial statements based on the earth mover's distance (EMD) (Rubner et al., 2000). Our approach takes into account both the structural properties as well as the value-related information of financial statements. We argue that both the hierarchical structure of the chart of accounts and the value distribution across the ledger accounts should be considered in order to accurately identify company peers. This will ensure that companies with similar financial structures and value distributions are identified as similar, while those with very different structures or value distributions are not considered similar.

1.5 | Contributions

Our main contributions are summarized as follows:

- We introduce a graph-based company distance metric that takes into account both structure and value information.
- We offer a detailed description of how our graph distance metric can be applied to the balance sheet component of a financial statement, but this work is by no means limited to financial applications alone. Users can customize the distance metric to their specific needs, resulting in a user-tailored peer selection method.
- We provide an efficient implementation of the proposed approach, making it practical to use the distance metric on a large scale.
- We conduct an experimental study using real-world financial data to demonstrate the usefulness of our proposed distance metric. Through our analysis, we show that our method outperforms the state-of-the-art in identifying peer companies that are financially related and engage in similar business activities.
- We propose a low-dimensional company representation based on financial statement information. This company embedding is transferable and adaptable for downstream prediction tasks, and to the best of our knowledge, it is the first of its kind to be based on financial statement information. Additionally, we show that the low-dimensional company representation based on our proposed distance metric demonstrates superior performance, capturing more industry information than existing methods.

In conclusion, our distance metric has the potential to be useful for a variety of applications, including helping investors identify investment opportunities, aiding government officials in detecting fraudulent behavior, and allowing accountants to benchmark a group of companies.

The remainder of the paper is structured as follows. In Section 2, we provide a summary of the related work in the field. Section 3 introduces the graph distance metric for financial statements. The experimental evaluation of our method can be found in Section 4. Finally, in Section 5, we conclude this work and provide suggestions for future studies.

2 | RELATED WORK

In this section, we review related work in the field of company peer selection.

2.1 | Industry classification systems for company peer selection

In general, identifying similar companies is a demanding and time-consuming task that depends on manual skill and professional knowledge. As a result, several attempts have already been made to streamline this process. Currently, the task of selecting similar peers relies mostly upon industry classification systems, an economic taxonomy that organizes companies into economically-related groups. Industry classification systems allow companies to be classified into homogeneous categories with the assumption that companies within the same group display similar characteristics. The Statistical Classification of Economic Activities in the European Community (NACE) is the classification standard of the European Union. This classification standard can be compared with the North American Industry Classification System (NAICS), used in the United States, Canada, and Mexico. It is well recognized that industry classification aids in company analysis when compared to simply considering company size (Kahle & Walkling, 1996). The use of industry codes to create peer company groups has several advantages: It is simple to understand and simple to implement.

Industry classification approaches, however, have their limits. One disadvantage is that classification systems do not evolve at the same rate as the market conditions, making new industries difficult to classify (Fan & Lang, 2000). Another disadvantage is the lack of consistent classification standards, which results in varying company classifications depending on the industry classification standard chosen (Yang et al., 2019). Ultimately, it is unlikely that a company's peers will remain unchanged for an extended period of time. That is why rapidly altering business environments necessitate a more flexible classification system to keep up with the rapid transformation of the market (Ding et al., 2019). This implies that we cannot merely distinguish organizations based on their industry classification. Companies provide a broad range of products and services and assigning them to a single category, would oversimplify the company's business. Numerous companies are assigned to one category but the differences between those companies are too big to disregard.

Instead of categorizing a company, we advocate for a system that allows us to compute a similarity score based on company data. This enables us to quantify the resemblance between two companies regardless of their industry activity codes. By measuring the distance between two companies, we will be able to identify peer companies with much finer granularity. Numerical similarity measurement is especially important when only a small number of peer companies need to be selected from the group consisting of many companies with diverse profiles (Kee, 2019).

2.2 | Company peer selection approaches and company distance metrics based on text and non-financial data sources

Academia has been relatively slow to create innovative, goal-oriented peer selection methods, instead of relying solely on industry classification schemes (Ding et al., 2019). Despite this, various attempts have been made to identify similar companies. In examining the relatedness between companies, research studies have been carried out on a variety of data sources. Fan and Lang (2000) employed input–output commodity flow data and introduced two IO-based measures to capture company relatedness. Another study makes use of patent information to form peer groups based on technological proximity (Gkotsis et al., 2018). Asche and Misund (2016) employ econometric techniques to explore the usefulness of company valuation multiples in identifying similar companies.

The use of company textual data has also been investigated to quantify company relatedness. Several studies make use of the textual data present in annual 10-K filing reports. Hoberg and Phillips (2016) make use of the product descriptions present in the company filings. They compute the pairwise word similarities between the bag of word representations of the product descriptions. Fang et al. (2013) and Shi et al. (2016) make use of the business description present within the 10-k filing reports. They both apply latent Dirichlet allocation (LDA) to the business descriptions to discover topic features that indicate business commonalities between companies.

Aside from filing reports, company website information and news articles have been utilized to identify company peers. Lee et al. (2015) state that companies appearing in historically adjacent searches by the same person are similar. Berardi et al. (2015) extract text-related features from company websites to build a model that classifies companies by industry sector. Bernstein et al. (2003) observe the co-occurrence of companies in news articles and introduce a relational vector-space model that abstracts the linked structure, representing companies by weight vectors. Finally, Kee (2019) utilized Word2Vec to obtain word embeddings from a 10-year collection of news articles about companies. The pairwise cosine similarity of the word embeddings was used to identify company peers.

Finally, Raman et al. (2019) employ a company network as a data source, with the edges between the companies representing their relationship. They employ a graph representation learning method that allows them to create a company embedding that considers the interrelationship between the companies.

As research shows, various data representations and information sources have been used to obtain company peers. However, an individual's personal objectives may influence their choice of peer selection method, as different strategies may be more beneficial depending on the situation. It is important to note that no single peer selection strategy is universally applicable, as different approaches may be more suitable in different situations (Ding et al., 2019).

2.3 | Peer selection and company distance metrics based on financial statements

A different approach to identifying company peers involves using company financial statements to measure their relatedness.

2.3.1 | Financial statements

In accounting and financial reporting, companies have several accounts that together compose a hierarchy described by the chart of accounts. In short, it is an organizational framework that provides a digestible breakdown of all the financial transactions that a company conducted during a specific accounting period. Broken down into subcategories included in the financial statements of a company, this information is presented in a structured manner in which it is easy to understand. Financial statements typically include balance sheets, statements of profit or loss, and reconciliations. In addition to providing a concise and comprehensive description of a company's financial condition and operational performance (Bushman & Smith, 2001), financial statements fairly portray the business activities from a financial standpoint.

Additionally, a financial statement accurately reflects the relationship between assets, liabilities, expenses, and revenue structures which is essential for understanding the financial situation of a company (Yang & Cogill, 2013). This is why we prefer an automated method for identifying common financial disclosure structural and value patterns, which can assist information users in locating similar

peers based on financial statement information that inherently captures their common business operations and strategies.

2.3.2 | The benefit of comparing financial statements

Finding similar financial statements is appealing for a variety of reasons. Investors believe financial statements to be a reliable predictor of business performance. Hopkins (1996) claims that the inspection of the balance sheet has an impact on financial experts' assessments of the stock price. This implies that companies with comparable financial sheets ought to perform similarly if successful companies are identified. Furthermore, a number of studies have argued that companies with similar business activities ought to have comparable financial statements (De Franco et al., 2011; Yang & Cogill, 2013). Yang and Cogill (2013) present evidence that a company distance metric may successfully identify industry borders, which supports this viewpoint.

Aside from the similarity of financial statements, dissimilar financial statements can also provide useful information. Fraud detection is one example where this is shown (Jan, 2018; Kanapickienė & Grundienė, 2015). Companies may alter their financial data to increase stock prices or get access to long-term debt funding. Therefore, it is quite useful to have a company distance measure that allows one to quantify how dissimilar two companies are. This enables regulatory authorities to recognize these unusual financial statements. Deviations may also highlight certain company characteristics. These distinctive qualities could suggest that a company is in a unique position, which might suggest a promising investment opportunity (Yang & Cogill, 2013).

2.3.3 | Approaches using financial ratios

Several attempts have been made to determine the relatedness of companies based on their financial statement information. Financial ratios are one method for determining how similar two companies are (Kanapickienė & Grundienė, 2015). Financial ratios rely on data extracted from financial statements to provide meaningful numerical values that indicate a company's current operating activities or financial performance. Ding et al. (2019) use K-medians clustering to find peer companies by using a set of financial ratios. This means that a set of financial ratios should be selected in order to compare the financial performance of companies. As a result, selection bias may be introduced into the process. Another limitation is that corporations may use window dressing to improve their financial figures. Financial analysts must be careful of activities that artificially inflate a company's solvency or liquidity.

2.3.4 | Approaches using the full financial statements

Another line of research tries to address this issue by looking at the financial statements as a whole. Brown et al. (2021) represent a company as a vector where each element represents a ledger account value. They define the similarity between two companies as the cosine or Mahalanobis distance between these vectors. This results in a numerical value expressing how similar two companies are. Hoitash et al. (2018) proposed another technique in which a company is represented by the set of ledger accounts that appear in their financial statement. As a result, using the Jaccard index, they define company relatedness as the overlap in ledger accounts between those two companies. These techniques, however, do not take into account the structural properties of a financial statement. More specifically, the relatedness and hierarchical position of ledger accounts within a financial statement have no effect on the distance measure. This means that two closely related ledger accounts, such as *agricultural land* and *residential land*, have the same influence on the distance measure as two ledger accounts that are completely unrelated.

The paper of Yang and Cogill (2013), advocates for exploiting the structural properties of the ledger accounts found in a financial statement. They created a tree edit distance-based algorithm that considers companies to be similar if their financial statement structures are similar. This technique only evaluates the structure of a company's financial statement. This means that the ledger account values are not taken into account. This inspires the idea of considering both the financial statement structure as well as the values on the ledger accounts when determining the similarity between two companies.

Nonetheless, financial statements as a whole are not commonly used to compare companies. The automated use of financial statement information has been rather limited as a result of the lack of standardization and frequently depends on laborious human processing. A standardized method for digitizing financial statements is required to increase the efficiency of information processing by allowing the financial data of various businesses, sectors, and reporting periods to be normalized for automated analysis. With the advancement of information technology, there is an increased interest in utilizing technology to improve information processing speed (Cong et al., 2014). This emphasizes the need for a company distance metric that can quantify company similarity in a data-driven fashion.

3 | GRAPH DISTANCE METRIC

In this section, we will first introduce the tree representation of a financial statement and explain the rationale behind our proposed distance metric. We will then provide a visual representation of our method and describe the mathematical details of both a naive and efficient implementation of the distance metric. Finally, we will compare the computational cost of both implementations and discuss the rationale behind the weight function used.

3.1 | Financial statements as a graph

Hierarchical structures with additional semantic links and cross-references can be used to depict financial statements. As established by Yang and Cogill (2013), a vertex-labeled tree is a natural

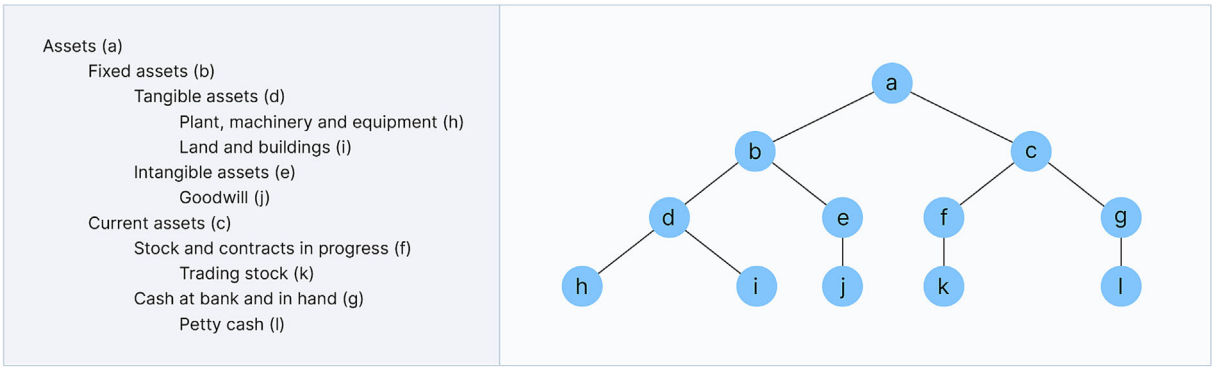


FIGURE 1 Left: *Assets* subsection of the balance sheet. Right: A vertex-labeled tree representation of the *assets* subsection of the balance sheet.

representation of the ledger accounts included in a financial statement. In other words, vertices have distinct account names assigned to them, and an ordered tree structure could be utilized to explain the semantic relationships between these financial concepts.

We use the *assets* part of a balance sheet as an example. There are two categories of *assets*: *fixed assets* and *current assets*. As a result, a ledger account may be partitioned into other, more specific accounts. Figure 1 illustrates how the ledger account for *plant, machinery, and equipment* is part of the category of *fixed assets*, which is further separated into *tangible assets* and *intangible assets*. Ledger accounts can also be included in the *current assets* category, which is further separated into *stocks and contracts in progress* and *cash at bank and in hand*. It should be noted that the vertex-labeled form of a balance sheet is not exclusive to this particular illustration. For illustration purposes, a subset of ledger accounts and their relationships are provided.

Clearly, the internal structure of a financial statement is preserved by this representation method. This structural format can also be used to display profit and loss statements, in addition to balance sheets. As a result, a vertex-labeled tree with ledger account names as vertex labels may be used to depict a company's financial statements. The tree of all possible financial accounts hierarchically organized inside a financial statement serves as the general structure of a financial statement, allowing us to represent any organization.

In this study, a company's financial statement is represented as a vertex-weighted tree. This implies that a weight is assigned to each node in the tree. This weight is given to a particular node based on the value of its ledger account; in particular, it equals the node's relative importance, as defined by the weight function w mentioned in Section 3.3.

The automated use of financial statement information has been very limited due to a lack of standardization, which often relies on laborious manual processing. A uniform method for digitizing financial statements is needed to increase the efficiency of information processing. This would allow the financial data of various companies, sectors, and reporting periods to be standardized and subject to automated analysis. Because of Silverfin's automated approach that maps financial statement data onto a standardized chart of accounts, the obstacle for comparing financial statement information becomes significantly smaller.[1]

3.2 | Rationale behind the proposed distance metric

First and foremost, it is critical that our distance metric can take into account the hierarchical structure of the ledger accounts found within a financial statement. This signifies that the distance metric may identify comparable companies based on the structure of their financial statements. In other words, the metric must be able to understand and take into account the similarity of ledger accounts and how they hierarchically compose the chart of accounts. Assume that the ledger account for *land and buildings* can be further subdivided into *agricultural land* and *residential land*. These two nodes should be considered more similar by the distance metric than two ledger accounts that are located further apart in the graph. In addition, it is also important that the way the ledger accounts hierarchically compose the chart of accounts is taken into account. Companies that structure their ledger accounts in a similar way should be considered more similar. The same goes for the reverse story, if two companies are made up from very different and not similar ledger accounts, the distance between these two companies should be large. In this paper, we refer to **structure** as the hierarchical composition of the different ledger accounts present within a company's financial statement.

Although companies can have a very similar financial structure, it is of course equally important to take into account the values situated on their different ledger accounts. Companies that have a very similar financial statement structure but a very different value distribution over their ledger accounts should not be regarded as similar. Think about a scenario where two companies share a very similar balance sheet structure. For one company, the largest ledger account weight could be located on the *buildings* node while for the other company, the largest weight could be located on the *trading stock* node. For this reason, it is also important that the value distribution across the

different ledger accounts is taken into account in order to identify similar companies. In this paper, we refer to *value* as the values situated on the ledger accounts within a financial statement.

For the reasons stated above, we suggest a distance metric that considers both the structure and values of a financial statement. Furthermore, we believe that methods that accomplish this are more effective at discovering similar companies than methods that simply analyze one of the two sources of information.

3.3 | The earth mover's distance based graph distance metric: EMD-GraD

A generic tree can be used to represent the financial information of each company. The generic tree is denoted by $T=(V,E)$, where V is a collection of $|V|=n$ nodes, and E is the collection of edges. Next, we define a company-specific weight function $w: V \mapsto \mathbb{R}$, which converts the generic tree into a company-specific tree by assigning a weight to each node. The generic tree T and the weight function w allow us to map every company's financial statement into a company-specific tree representation.

Consider $T_1=(V,E,w_1)$ as the tree representation for company 1, and $T_2=(V,E,w_2)$ as the representation for company 2. Based on our motivation, we propose a distance metric that considers the structural properties as well as the node weights of T_1 and T_2. We define the distance between two vertex-weighted trees as the total cost of moving weights over the edges of T_1 in order to become identical to T_2. A company that is slightly different from another company in terms of balance sheet structure and ledger account value distribution does not necessitate numerous weight transfers. On the other hand, significantly different companies necessitate many weight transfers.

This distance metric is based on the earth mover's distance (Rubner et al., 2000). According to EMD literature, tree 1 acts as the source tree, and tree 2 as the sink tree. It is important to note that this distance is symmetric. Figure 2 depicts a graphical representation of how the distance metric works. This is formally defined in Definition 1.

Definition 1 (Earth mover's distance based graph distance metric. Given an undirected graph $T=(V,E)$ with $|V|=n$ and two weight functions $w_1: V \mapsto \mathbb{R}$ and $w_2: V \mapsto \mathbb{R}$ where $\sum_{i=1}^{n} w_1(v_i) = \sum_{i=1}^{n} w_2(v_i)$. Consider $T_1 = (V,E,w_1)$ as the source graph where $p_i = w_1(v_i)$ is the production weight associated with node i, also consider $T_2=(V,E,w_2)$ as the sink graph where $c_i = w_2(v_i)$ is the consumption weight associated with node i. Then the distance between the graphs T_1 and T_2, denoted as $\phi(T,w_1,w_2)$, is defined as the minimum amount of total weight allocation that has to be shifted over the edges of T_1 in order to become identical to T_2.

To compute this distance, a linear programming problem can be solved to find the edge flows $f_{i\to j}$ (with $f_{i\to j} \triangleq -f_{j\to i}$) that minimize the overall cost:

$$\text{minimize } C = \sum_{(i,j)\in E} |f_{i\to j}|,$$

subject to the constraint that, for every node i, the following must hold:

$$\sum_{j:(i,j)\in E} f_{i\to j} = p_i - c_i,$$

where the total flow for a node i is equal to the production p_i minus the consumption c_i.

This distance metric searches for the optimal flow matrix $\mathbf{F} \in \mathbb{R}^{n\times n}$ where the total distance is defined as $\sum_{i=1}^{n}\sum_{j=1}^{n}|\mathbf{F}_{ij}|$. Although the absolute value is a non-linear function, the optimization problem can be transformed into a linear program by introducing additional variables g_{ij} and constraints, as follows (Boyd et al., 2004):

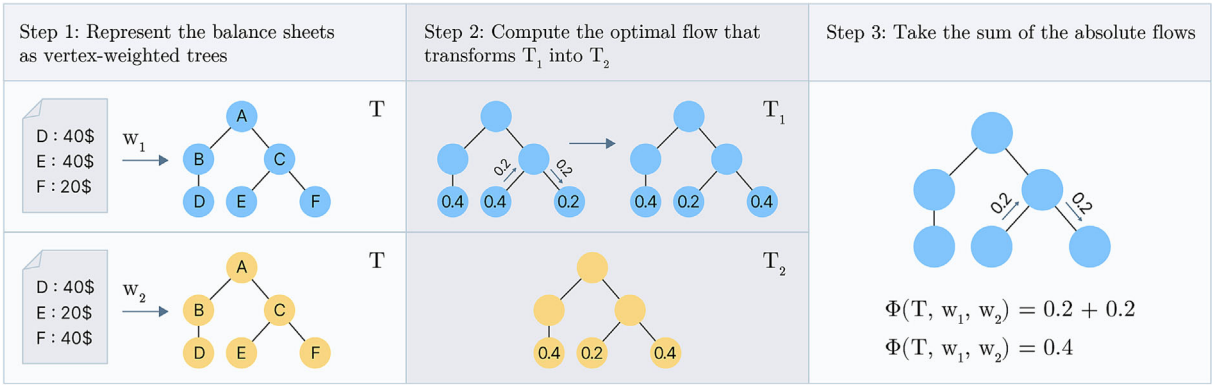


FIGURE 2 Graphical representation of how our proposed distance metric calculates the distance between two companies. Step 1 shows how the company-specific weight function w transforms the general tree structure T into a company-specific vertex-weighted tree. Step 2 computes the optimal edge-flows so that T_1 and T_2 become identical. Step 3 takes the absolute sum of the optimal edge-flows which represents the distance between two companies.

$$\text{minimize } C = \sum_{(i,j) \in E} g_{ij},$$

subject to the constraints:

$$g_{ij} \geq f_{i \to j},$$

$$g_{ij} \geq -f_{i \to j},$$

$$\sum_{j:(i,j) \in E} f_{i \to j} = p_i - c_i.$$

A naive implementation by solving this linear program can be very computationally intensive, which would limit the scalability of the distance metric. Fortunately, this linear programming problem has a special structure known as a network flow problem, which allows for efficient solution methods. By exploiting the structure of this problem, we propose an efficient implementation for computing the distance between two vertex-weighted graphs with a lower computational cost, making it suitable for use on large-scale graphs. In this section, we will first explain the efficient implementation of EMD-GraD using a vertex-weighted tree, $\Delta T = (V, E, w_1 - w_2)$, which represents the difference between the graphs T_1 and T_2. This implementation is defined by Algorithm 1. After that, we will also prove that this implementation is equivalent to the naive LP approach and provide the computational cost of the efficient implementation.

Algorithm 1 Efficient Implementation

Input: ΔT(generic vertex-weighted tree)
Output: tree_distance

```
function ComputeTreeDistance(ΔT)
    tree_distance ← 0
    leaf_nodes ← GetLeafNodes(ΔT)
    while leaf_nodes != [root node] do
        for v ∈ leaf_nodes do
            tree_distance ← tree_distance + |v.weight|
            v.parent.weight ← v.parent.weight + v.weight
            remove v from ΔT
            if v.parent.children == [] then
                leaf_nodes ← leaf_nodes + [v.parent]
    return tree_distance
```

The algorithm above computes the distance between the graphs T_1 and T_2 by looking at the leaf nodes of the tree ΔT. For each leaf node, it takes the absolute value of the weight of that node and adds it to a running total called 'tree_distance'. Then it takes the weight of the leaf node and adds it to the weight of the parent node of that leaf node, and removes the leaf node from the tree. This is done for all leaf nodes in the tree until there are no more leaf nodes left. The final value of 'tree_distance' represents the distance between the two graphs being compared.

Prior to demonstrating the equivalence of this implementation to the naive LP approach (see Theorem 1), we first introduce Proposition 1. To fully grasp Proposition 1, we also introduce a new weight function, $w_0(v_i)$, which assigns a weight of 0 to each vertex v_i.

Proposition 1. *The distance between the graphs T_1 and T_2, denoted as $\phi(T, w_1, w_2)$ can be reformulated as $\phi(T, w_1 - w_2, w_0)$.*

Proof. Let C be the minimum cost of the linear programming problem defined by the objective function

$$\text{minimize } C = \sum_{(i,j) \in E} |f_{i \to j}|$$

and the constraint that for every node i,

$$\sum_{j:(i,j) \in E} f_{i \to j} = p_i - c_i.$$

Then the distance between the graphs T_1 and T_2, denoted as $\phi(T, w_1, w_2)$, is equal to the minimum cost C.

Now, consider the zero weight function $w_0(v_i)$ that maps the weight of each vertex v_i to 0. Then the distance between the graphs T_1 and T_2 as defined by the new weight functions $w_1 - w_2$ and w_0 is equal to the minimum cost of the linear programming problem defined by the objective function

$$\text{minimize } C = \sum_{(i,j) \in E} |f_{i \to j}|$$

and the constraint that for every node i,

$$\sum_{j:(i,j) \in E} f_{i \to j} = (w_1 - w_2)(v_i) - w_0(v_i).$$

Since $w_0(v_i)$ is equal to 0 for every vertex v_i, we have that

$$\sum_{j:(i,j) \in E} f_{i \to j} = (w_1 - w_2)(v_i) - w_0(v_i)$$

is equal to

$$\sum_{j:(i,j) \in E} f_{i \to j} = (w_1 - w_2)(v_i).$$

Therefore, the distance between the graphs T_1 and T_2 as defined by the new weight functions $w_1 - w_2$ and w_0 is equal to the distance between the graphs T_1 and T_2, denoted as $\phi(T, w_1, w_2)$. □

In the next theorem, we show that the efficient implementation described in Algorithm 1 correctly computes the distance between

two vertex-weighted graphs using the naive LP approach of EMD-GraD.

Theorem 1. *The distance between the graphs T_1 and T_2, denoted as $\phi(T,w_1,w_2)$, can be computed by the efficient implementation described in Algorithm 1.*

Proof. As per Proposition 1, the distance between the graphs T_1 and T_2 can be reformulated as $\phi(T,w_1-w_2,w_0)$, where w_0 is the zero function.

We will prove this theorem by showing that the efficient implementation described in Algorithm 1 can find a solution that minimizes the overall cost of $\phi(T,w_1-w_2,w_0)$.

Consider a leaf node i and let p be its parent node. Since $\Delta T=(V,E,w_1-w_2)$ is a tree, the edge (i,p) is the only edge incident to node i. Thus, the constraint that for every node i,

$$\sum_{j:(i,j)\in E} f_{i\to j}=f_{i\to p}=(w_1-w_2)(v_i)$$

implies that $f_{i\to p}$ is equal to the weight of node i. Therefore, sending all the weight of node i to its parent node p is the unique solution that minimizes the overall cost.

If this is the case for all leaf nodes, we can prove through induction that sending all the weights from the children to their parents minimizes the objective function of the LP problem. Algorithm 1 follows this method by iterating through all the leaf nodes and sending their weight to their parent nodes. Since this is a valid solution that minimizes the overall cost, we can conclude that the distance between the graphs T_1 and T_2, $\phi(T,w_1,w_2)$, can be computed by the efficient implementation described in Algorithm 1. □

The efficient implementation drastically decreases the computational cost of the distance metric. One drawback is that by simplifying the calculation of the distance metric, interesting information is omitted. For example, the optimal flow matrix **F** is abstracted, which means that there are no more traceable flows when calculating the distance. This prevents the addition of an explainability layer that would be able to explain to what extent two companies differ the most.

In the next section, we compare the computational cost of both implementations and show that, if such explainability is not required, it is better to use the efficient implementation.

3.4 | The computational cost

It has been proven that a general quadratic program (QP) is NP-hard (Vavasis, 1991). Because our QP problem can be translated into a constrained linear program, polynomial time algorithms can be used to compute the distance metric. Vaidya (1989) presented an algorithm for solving linear programming problems that require $\mathcal{O}((m+n)^{1.5}nL)$ arithmetic operations, where m is the number of constraints, n the number of variables, and L a parameter defined in the paper. State-of-the-art algorithms speed-up the time complexity to $\tilde{\mathcal{O}}(n^{2+1/6}L)$ and $\tilde{\mathcal{O}}(n^{2+1/18}L)$ (Jiang et al., 2020).

The efficient implementation of EMD-GraD significantly improves the time complexity of the naive LP approach. This makes it more interesting to use the distance metric on a large scale. To demonstrate the advantage of the efficient implementation, we prove the time complexity of EMD-GraD.

Theorem 2. *The computational cost of the efficient implementation (Algorithm 1) for computing the earth mover's distance based graph distance metric is $\mathcal{O}(n)$, where n is the number of nodes in the tree.*

Proof. The computational cost of the algorithm is $\mathcal{O}(n)$ because it iterates over all n nodes of the tree once and performs a constant number of operations on each node. The algorithm starts by retrieving the leaf nodes of the tree, for example, through depth-first traversal, which has a computational cost of $\mathcal{O}(n)$. Thus, we have

$$C=\mathcal{O}(n).$$

□

In addition to the proof, Figure 3 displays the computational time taken by both implementations. This figure shows how the computing times of the two implementations vastly diverge as the number of nodes in the perfect binary tree rises, with the efficient implementation performing noticeably better.

3.5 | The weight function

In this section, we go into further depth about how generic tree representation T and the weight function w were determined.

3.5.1 | The generic tree representation

Instead of considering the distance between two companies as the distance between the different parts that make up a balance sheet, we include the debit active, credit active, debit passive, and credit passive into a single generic graph. By seeing a financial statement as a forest of many trees, the method avoids the potential of shifting weights between the active and passive sides of a balance sheet. Understanding the interaction between the assets, liabilities, expenses, and revenue structure is directly related to understanding a company's financial position (Yang & Cogill, 2013).

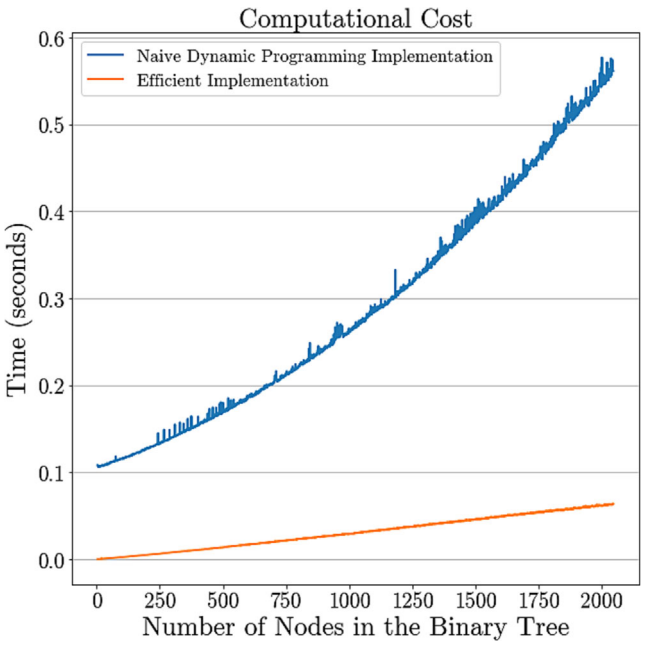


FIGURE 3 The time taken by the two different implementations with variable tree sizes.

3.5.2 | Determining the weight function

The weight function we use in this paper can be represented by the following function $w(v_i) = \frac{b_i}{\sum_{i=1}^{n} b_i}$, where b_i represents the value of ledger account node i. The weight assigned to a ledger account reflects the relative importance of a node within the tree. This makes it simple to understand the weight of a node: node weights where $w_1(v_i) > w_2(v_i)$ represent the situation where the ledger account i is more important for company one compared to company two, node weights where $w_1(v_i) < w_2(v_i)$ represent the opposite, and node weights where $w_1(v_i) = w_2(v_i)$ represent the situation where both companies consider the ledger account i as equally important.

Within a balance sheet, negative ledger account values are present. On the active side of a balance sheet, a negative ledger account value denotes a credit account (e.g., a depreciation), while a positive value on the passive side denotes a debit account. When both debit and credit ledger account values are included in the active or passive tree, the concept of explainability is constrained, making interpretation of the weights more difficult. However, considering a generic tree that takes into account every aspect of the balance sheet improves the performance of the distance metric. Although it is not a focus of this study, more research can be done to determine the importance of a node by analyzing how the distinct flow costs contribute to the absolute cost between two graphs.

3.5.3 | A user-tailored peer selection method

Finally, we would like to emphasize that the weight function w is adaptable to specific circumstances. To compute the node weights, other custom functions as well as different node attributes can be used. A company's feature vector **b**, which is a vector of all booked values in the balance sheet, can be easily replaced by another feature vector. Instead of using the actually booked values, another option is to use the number of transactions associated with a certain ledger account. The only prerequisite is that $\sum_{i=1}^{n} w_1(v_i) = \sum_{i=1}^{n} w_2(v_i)$ when calculating the distance between two vertex-weighted graphs. This allows users of this distance metric to create their own version of the metric that is tailored to their specific needs.

4 | EMPIRICAL EVALUATION

We conduct various experiments on real-world financial data to verify the usefulness and efficacy of the distance metric. This enables us to determine whether it could be advantageous for someone looking for peer companies. In this section, we perform two experiments where our method is compared to other benchmark methods. Each empirical evaluation is preceded by a description of the experimental setup and is concluded by an interpretation of the experiments.

During the empirical evaluation, we want to answer several questions.

- Does a method that takes into account structure and value information provide more information than methods that consider one of the two in the process of peer selection?
- Does the proposed distance metric allow one to construct a company embedding that contains meaningful information which can be used for follow-up tasks?
- Does the new method allow one to find similar companies?

This section starts by describing the data, followed by the introduction of other peer selection methods used as baselines for comparison. Afterward, we present two distinct experiments where we will extensively examine the value of the new method based on real-world data.

4.1 | Data

We use proprietary data from Silverfin,[2] a Belgian scale-up focused on building an accountancy cloud service. For the purpose of building our SILVERFIN dataset, we obtain the private financial statement data from Silverfin, along with the financial ratios and industry activity codes assigned to a company. The dataset is composed of 2839 companies spanning a wide range of industrial sectors, geographical regions, and market sizes. Since this is real-world financial data, accountants using Silverin's service may be able to gain significant insights. Sample data[3] are made available on GitHub in order to evaluate the distance metric because the data

used to conduct the experiments is confidential and cannot be shared.

For every company, we have the financial statement information, organized by a standardized taxonomy maintained by Silverfin. Additionally, we have a generic tree with all possible ledger accounts that are used by the companies in the dataset. This enables us to generate the vertex-labeled graph representations of the various companies by mapping the financial statement data onto the generic tree.

In the experimental setting, we focus on 2839 Belgian companies whose financial data was derived from their fiscal year 2019 financial statements. Table 1 presents the data summary of the sample companies from the year 2019. It also contains the overall industry activity information over the sample companies. In addition to identifying the various NACE codes used by the sample companies, we determine the number of producible groups when the first two or three NACE code numbers are considered.

The distribution of industry codes among the various companies can be seen in Table 2, where each company has between 1 and 20 different industry codes. Because industry classification systems merely focus on summarizing all the performed business activities within a company, determining the most important business activities and hence industry activity codes is difficult.

To evaluate the performance of EMD-GraD, we validate our method against several financial ratios, considering them as ground truth information. In Table 3, we have provided a list of financial ratios that have been included in our analysis. For each ratio, we have included a definition, which summarizes all the necessary information needed to compute the ratio. To compute the financial ratios per company, the relevant financial information was extracted from a company's financial statement or other information sources present in the Silverfin database.

Table 4 provides the descriptive statistics of the financial ratios used to validate the proposed distance metric. Before the financial ratios were computed we removed the companies that did not have the necessary information to compute the ratios. Additionally, in order to reduce the outlier effect, we eliminated the companies with financial ratios that fall outside the 5th and 95th percentiles.

TABLE 1 Data distribution summary.

Fiscal year	2019
Number of sample companies	2839
Number of different NACE codes used by the sample companies	622
Groups using two-digit NACE industry codes	79
Groups using three-digit NACE industry codes	215

TABLE 2 NACE code distribution.

	Number of NACE industry codes per company
Mean	2.62
Std.	1.71
25th percentile	1
Median	2
75th percentile	3
Max	20

TABLE 3 Financial ratio definitions.

Ratios	Definition
Debt Ratio	$\frac{\text{Total debt}}{\text{Total assets}}$
Current Ratio	$\frac{\text{Current assets}}{\text{Current liabilities}}$
Quick Ratio	$\frac{\text{Current assets - Inventory}}{\text{Current liabilities}}$
Days Client Credit	$365 * \frac{\text{Trade debtors}}{\text{Turnover} + \text{Other operating income}}$
Value Added over Tangible Fixed Assets	$\frac{\text{Value added}}{\text{Tangible fixed assets}}$

4.2 | Methods

Because we are unaware of other methods that take into account both the structure and value information present in financial statements, we compare our method against methods that take into account one of the two.

We compare our approach against the following methods as baselines:

- **Tree edit distance**: Yang and Cogill (2013) developed an algorithm based on the tree edit distance that can effectively identify industry boundaries by taking into account the structural-semantic information present within financial statements. In their paper, they translate the underlying company graphs into property strings and use the Levenshtein (Levenshtein, 1966) distance to compute the pairwise similarity.

TABLE 4 Financial ratio data distribution.

Ratios	Obs.	Mean	Std.	25th percentile	Median	75th percentile
Debt Ratio	2548	0.62	0.31	0.38	0.63	0.90
Current Ratio	2538	2.18	2.48	0.82	1.32	2.45
Quick Ratio	2538	1.93	2.40	0.54	1.11	2.18
Days Client Credit	2595	70.45	101.04	6.54	34.69	79.34
Value Added over Tangible Fixed Assets	2151	1.35	3.23	0.08	0.31	1.15

- **Ledger account overlap**: This method compares two companies by computing the percentage of overlapping financial statement items, ignoring ledger account value information. The pairwise distance is obtained by computing the Jaccard similarity between two sets of ledger accounts present within each company's financial statement (Hoitash et al., 2018).
- **Cosine similarity score**: Brown et al. (2021) represent a company as a vector where each element represents a ledger account value. The pairwise distance is computed by computing the cosine similarity between two company vectors. This means that only the value information is considered.
- **Mahalanobis similarity score**: This method, introduced by Brown et al. (2021), starts from the company ledger account value vectors and computes the pairwise company distances by computing the Mahalanobis distance between the company vectors while ignoring the internal structure of a financial statement.
- **Random method**: This method randomly selects a set of companies and ranks them in a random way. This ranking introduces a degree of similarity based on the ranking of the companies.
- **EMD-GraD**: The earth mover's distance-based graph distance metric we propose in this paper (see Section 3.3).

4.3 | Experiment 1: finding peer companies

In this section, we focus on determining financially related peer companies that show similar characteristics in their business activities and in their financial performance. Along with testing the hypothesis that the selected peer companies are similar, we also examine whether including structure and value information improves the effectiveness of the distance metric. Using the different distance measures (see Section 4.2), we calculate the *k*-nearest neighbors of each company using the pairwise distances between the companies based on their financial statements.

Experiment 1 introduces two subexperiments. The first experiment examines which method best succeeds in establishing a peer group where we can find the highest concentration of companies performing the same activity. The second experiment examines which metric performs best for selecting financially related peers based on a set of financial ratios.

4.3.1 | Industry activity code analysis

Algorithm 2 explains how we calculated the industry activity overlap between the peer companies. The algorithm computes the average Jaccard similarity between the set of NACE codes of a company s_{nace} and their *k*-nearest neighbors. The computation is done for a set of *S* companies, after which the average is computed. The average Jaccard similarity represents how well the NACE codes of a company and their *k*-nearest neighbors overlap.

Algorithm 2 Industry Nearest Neighbors

Input: S(company set),k(number of neighbors), $\mathbf{D}$(distance matrix)
Output: $average\_jaccard\_score$

function ComputeJaccardScore($S, k, \mathbf{D}$)
  $jaccard\_score \leftarrow 0$
  for $s \in S$ **do**
    $predictions \leftarrow GetNearestNeighbors(s, k, \mathbf{D})$
    $s_{nace} \leftarrow GetNaceSet(s)$
    $z \leftarrow 0$
    for $p \in predictions$ **do**
      $p_{nace} \leftarrow GetNaceSet(p)$
      $jaccard \leftarrow Jaccard(s_{nace}, p_{nace})$
      $z \leftarrow z + jaccard$
    $jaccard\_score \leftarrow jaccard\_score + z/k$
  return $jaccard\_score/Size(S)$

Not all companies in the dataset can be part of the company set *S*, for which the peer companies are sought. This is due to the fact that a substantial fraction of the companies lack neighbors that engage in similar activities. To check whether a company can be a part of the company set *S*, we present two parameters. The number of companies that share at least one NACE code with the company being examined is specified by the parameter *q*. The minimal Jaccard similarity of all the companies with at least one shared NACE code is specified by the parameter *r*. For this experiment, we set $q = 15$ and $r = 0.4$. This results in approximately 1150 suitable companies.

Additionally, some activity codes are omitted because these activities are reported by many companies, but do not accurately reflect the key service that they provide. The NACE codes "70.220," "64.200," and "82.990," which refer to consulting businesses in the fields of business operations, holdings, and other business services, were disregarded for this experiment.

Figure 4 shows the predictive performance in industry code overlap for the different metrics. The *x*-axis represents the number of selected neighbors $k \in \langle 1,2,\ldots,15 \rangle$. The *y*-axis represents the average Jaccard overlap in industry codes across the generated peer groups. There is approximately a 2% industry code overlap between the neighbors when the random approach selects a set of nearest neighbors. All other distance metrics significantly outperform the random technique. EMD-GraD yields the highest similarity score. When it comes to selecting the first neighbor, the new metric performs nearly six times better than the random method. The ledger account overlap metric and the tree edit distance metric perform equally well starting from $k = 7$. The cosine similarity metric marginally outperforms these techniques. EMD-GraD continues to stand out from the other methods. We may conclude that EMD-GraD outperforms state-of-the-art methods for identifying the peer companies with the highest industry code overlap.

Given that there are more than 800 industry activity codes, choosing peer companies with NACE code overlap is a challenging task. Moreover, the majority of companies have a wide variety of NACE codes. As noted in the motivational section (see Section 3.2),

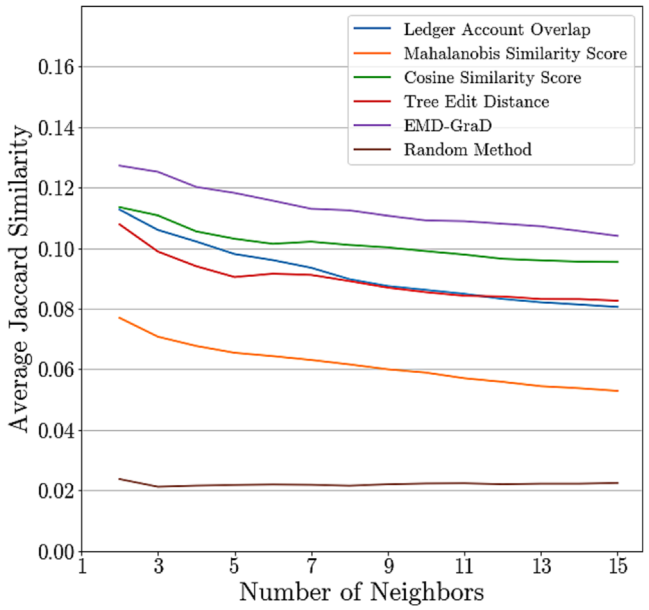


FIGURE 4 Average industry code Jaccard similarity between companies and their nearest neighbors.

we expect that companies could be very similar structure-wise, but very dissimilar based on their balance sheet distribution, and vice versa. This means that when considering both structure and value information, we anticipate a larger NACE code overlap among a company's nearest neighbors. We claim that this is the case because EMD-GraD beats those that merely take into account a portion of that information. This suggests that integrating both structure and value information enhances the usefulness of the distance metric.

4.3.2 | Financial ratio analysis

This experiment examines which metric performs best for selecting financially related peers based on a set of financial ratios. The variance of several financial ratios within a peer group is computed rather than the overlap in industry activity codes. In this case, the best strategy is one that creates peer groups with the lowest average variance in several financial ratios, as it is crucial to consider a company's financial status in addition to its performed activities.

Again this calculation is done for a set of S companies, with $q = 20$ and $r = 0.2$. The average variance of a financial ratio reflects how far the financial ratios of a peer group vary from one another.

Figure 5 shows the variance of the financial ratios within a peer group. The x-axis represents the number of k-nearest neighbors. The y-axis represents the average variance of the selected financial ratios across the peer groups. The financial ratios described in Section 4.1 are displayed in five different graphs. The random method, which does not consider any information, is generally the worst-performing one. There is one exception to this: In Figure 5e, the random method performs better than the Mahalanobis metric starting from nine neighbors. All other methods outperform the random method because they take financial statement information into consideration. However, the Mahalanobis similarity metric fails to adequately extract data from the financial records. EMD-GraD outperforms all other distance metrics on four out of the five ratios, although the margin of improvement varies. Our distance metric comes in second for the financial ratio "Value Added over Tangible Fixed Assets"; for this financial ratio, the overlap in the used ledger accounts appears to be more useful. For the ratios "Debt Ratio" and "Days Client Credit," EMD-GraD performs marginally better than the cosine similarity metric. EMD-GraD clearly outperforms every other baseline when it comes to the ratios 'Current Ratio' and 'Quick Ratio'. When it comes to finding financially related peers, the cosine similarity and the ledger account overlap method typically work well. The cosine similarity metric in particular appears to be useful for constructing peer groups.

Different distance metrics may preserve certain financial ratios better than others. Therefore, users should consider which metric best fits their purpose. As a result, a specific peer selection strategy may be beneficial in one situation but not in another, so it is important to choose the right strategy based on the specific situation (Ding et al., 2019). Overall, our method stands out from the competition because it more effectively considers financial performance information when selecting peers.

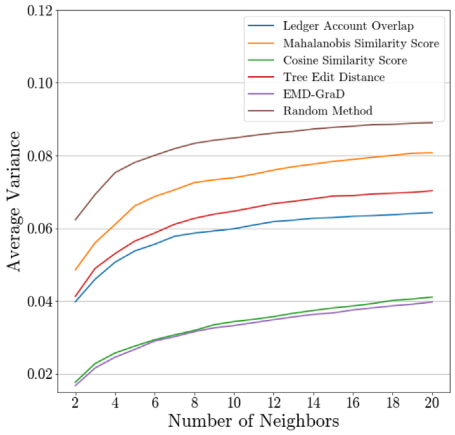
4.3.3 | Conclusion

We used the idea that companies can be compared based on how their ledger account values are distributed, but also on their financial statement structure. We have shown that approaches that consider both sources of information are more effective at locating peer companies. This experiment demonstrates that by taking into account both the structure of a financial statement and the values on the ledger accounts, we are better able to define the degree of similarity between companies, as EMD-GraD results in peer companies with a higher overlap in business activity and a lower variance in financial ratios.

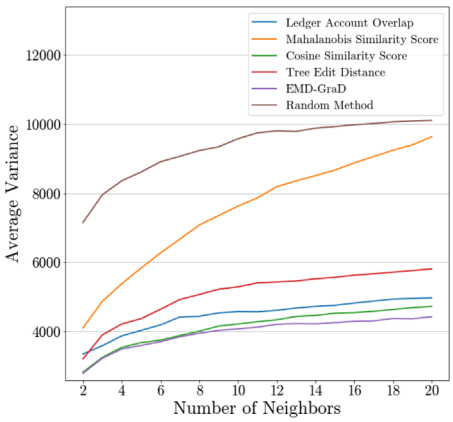
4.4 | Experiment 2: company embedding classification

In this experiment, we propose a financial statement-based classification system for industry codes that utilizes a low-dimensional representation of a company as input. By reducing the high-dimensional financial statement data into a more manageable representation, we are able to use it as input for downstream prediction tasks. The generated company representations are derived from the pairwise distances between companies, as calculated by the proposed EMD-based graph distance metric as well as the baseline financial statement distance metrics. These learned representations are transferable and can be easily adapted for a wide range of analytical tasks.

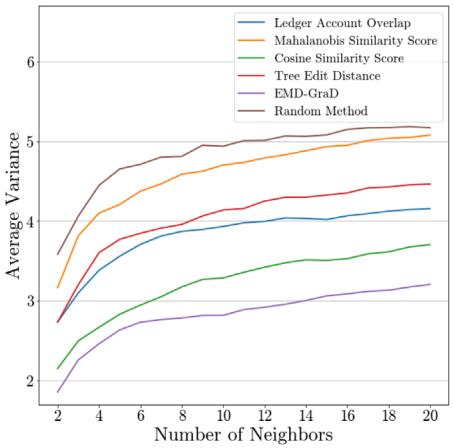
We hypothesize that EMD-GraD is superior to existing approaches for encoding companies into a low-dimensional space due to its consideration of both structural and value information. Additionally, we aim to evaluate the degree to which the company embeddings capture industry-related information by using them to predict industry codes.



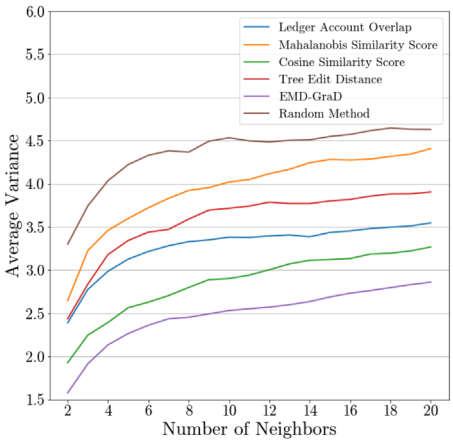
(a) Debt Ratio.



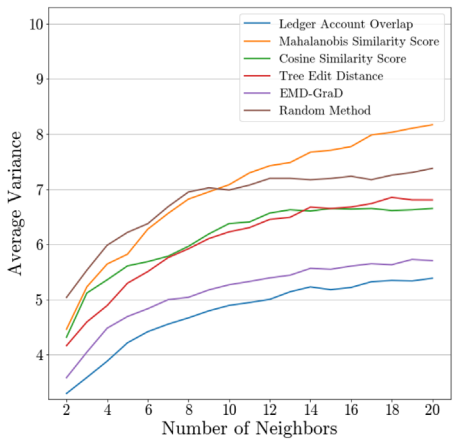
(b) Days Client Credit.



(c) Current Ratio.



(d) Quick Ratio.



(e) Value Added over Tangible Fixed Assets.

FIGURE 5 The average variance of five financial ratios among companies and their nearest neighbors.

4.4.1 | Experimental setup

In this experiment, we employ multi-dimensional scaling (MDS) (Kruskal, 1964) to generate low-dimensional representations of companies. MDS is a technique that involves projecting objects onto a lower-dimensional space in a way that maximally preserves the pairwise distances between them. By using MDS to reduce the dimensionality of the data, we aim to simplify the classification problem and improve the performance of our model. In this experiment, we explored the use of MDS with different dimensionalities. Specifically, we experimented with representing the data in MDS spaces of dimensions 5, 10, 15, 20, 25, 30, 35, and 40. A higher dimensional space

may provide more detailed information about the relationships between the companies, but may also be more difficult to interpret and process.

To assess and compare the quality of these low-dimensional representations, we trained 30 random forest models on every low-dimensional company representation and used different train-test splits in order to predict the industry codes of specific companies. A random forest is a popular machine-learning technique that uses multiple decision trees to make predictions (Breiman, 2001). By training multiple models on the same data with different train-test splits, we are able to assess the robustness of our predictions and evaluate the consistency of the results.

We state that the distance metric that allows us to obtain the highest overall accuracy levels, best captures the industry information within their low-dimensional representation. Additionally, we use a stratified model as a benchmark for evaluating the performance of our industry code prediction model. This approach is commonly used in machine learning as it allows us to compare the performance of our model to the performance of a model that respects the class distribution of the training data.

4.4.2 | The dataset

In this experiment, we group companies based on their primary activity by considering the first two or three digits of their NACE codes. This allows us to focus our analysis on specific industries or sectors. For example, if a company has been assigned several codes that begin with the two-digit prefix "86" (indicating the human healthcare sector), then the company will be assigned to this group only if all of its activity codes begin with these two digits. Otherwise, the company will not be included in the dataset.

We present four different use cases, as shown in Table 5, which summarizes the groups of companies, the number of samples in the dataset for each group, and the number of NACE code digits used for each use case. For instance, the first use case includes groups "86," "68," "62," and "01," with 88, 128, 37, and 34 samples, respectively, and aims to predict the correct company group based on their low-dimensional company representations.

To select the use cases for the two-digit industry classification groups, we randomly sampled nace code groups from the top 15 most frequently occurring groups. For the three-digit nace code groups, we also used a similar approach, but focused on ensuring a balanced representation of samples among all identified groups.

4.4.3 | Use case 1

The results of the first use case, as presented in Table 6, show the accuracies of the models trained on the reduced pairwise distances. The columns of the table correspond to the different dimension sizes obtained through MDS, while the rows represent the various distance metrics used. The "Mean" column provides the mean accuracy of each method across all dimension sizes, allowing for a general assessment of the performance of each method. The bold values highlight the best-performing method for each dimension size.

EMD-GraD is found to have the highest mean value, indicating that it is the most accurate method overall. This method also beats the others in six out of eight dimension sizes. The ledger account overlap and the cosine similarity metric also show good performance but are outperformed by EMD-GraD. The Mahalanobis similarity metric has the lowest mean value, suggesting that it this distance metric contains the least information. The stratified model achieves an accuracy of 0.396.

All methods are found to outperform the stratified model, indicating that there is valuable information in financial statements for predicting industry activity. EMD-GraD demonstrates the best overall performance and should be considered for future use in similarity analyses. This indicates that the company embedding obtained by using EMD-GraD contains the most valuable information.

TABLE 5 Summary of the executed experiments.

Use case	Group	Number of samples	NACE code digits
1	68, 86, 62, 01	128, 88, 37, 34	2
2	46, 43, 64	114, 90, 59	2
3	494, 433, 411, 561, 432, 960	31, 29, 26, 26, 23, 23	3
4	862, 682, 477, 642, 691, 683, 620	68, 58, 57, 57, 55, 37, 37	3

TABLE 6 Use case 1: accuracies in predicting the industry activity, obtained with the different distance metrics.

Methods	5	10	15	20	25	30	35	40	Mean
Ledger Account Overlap	0.594	**0.670**	0.631	0.633	0.622	0.633	0.626	0.628	0.630
Mahalanobis Similarity Score	0.446	0.403	0.378	0.407	0.385	0.407	0.385	0.413	0.403
Cosine Similarity Score	0.623	0.640	0.632	**0.662**	0.631	0.627	0.650	0.636	0.638
Tree Edit Distance	0.546	0.607	0.636	0.607	0.630	0.619	0.639	0.628	0.614
EMD-GraD	**0.632**	0.667	**0.652**	0.656	**0.654**	**0.641**	**0.658**	**0.646**	**0.651**

Note: The bold emphasis indicates the methods that achieve the highest accuracy for a specific vector dimension ranging from 5 to 40.

4.4.4 | Use case 2

The results of the models trained on the reduced pairwise distances for use case 2 are presented in Table 7.

The results in the table indicate that the different methods perform well in predicting industry activity based on financial statements compared to the stratified model that achieves an accuracy of 0.365. The ledger account overlap metric achieves the highest accuracy in one out of the eight dimension sizes, which is the same situation for the tree edit distance. EMD-GraD performs best in six out of the eight dimension sizes, achieving the highest mean accuracy overall. The Mahalanobis similarity metric has the lowest mean accuracy. Based on use case 2, EMD-GraD should be considered for future use in similarity analysis.

4.4.5 | Use case 3

For this use case, we try to predict the primary activity of companies considering their first three-digit NACE codes. The results of the experiment can be found in Table 8.

Based on the results presented in the table, it appears that EMD-GraD is the most effective at predicting the primary activity of companies based on their first three-digit NACE codes. The average performance of EMD-GraD is 0.410, which is higher than the average performance of the other methods. The next best performer is the cosine method, with an average performance of 0.388. The stratified model has an accuracy of 0.164, which is significantly lower than the accuracies obtained by the other methods. Overall, these results suggest that EMD-GraD is a promising approach for predicting the primary activity of companies based on their first three-digit NACE codes.

4.4.6 | Use case 4

The results of use case 4 can be found in Table 9. This use case considers the largest number of industry groups with the smallest amount of training data.

The results of the last use case show that EMD-GraD performs the best in terms of accuracy, with an average accuracy of 0.494 across all dimension sizes. This is followed by the cosine method, which has an average accuracy of 0.472. The ledger account overlap and the tree edit distance metric have similar performance, with average accuracies of 0.420 and 0.449, respectively. This is significantly higher than the performance of the stratified model that achieves an

TABLE 7 Use case 2: accuracies in predicting the industry activity, obtained with the different distance metrics.

Methods	5	10	15	20	25	30	35	40	Mean
Ledger Account Overlap	**0.586**	0.606	0.558	0.547	0.612	0.573	0.584	0.559	0.578
Mahalanobis Similarity Score	0.459	0.443	0.435	0.431	0.401	0.412	0.422	0.422	0.428
Cosine Similarity Score	0.530	0.602	0.585	0.601	0.626	0.594	0.579	0.607	0.590
Tree Edit Distance	0.523	0.602	**0.636**	0.623	0.638	0.592	0.610	0.644	0.608
EMD-GraD	0.582	**0.657**	0.635	**0.662**	**0.640**	**0.645**	**0.676**	**0.653**	**0.644**

Note: The bold emphasis indicates the methods that achieve the highest accuracy for a specific vector dimension ranging from 5 to 40.

TABLE 8 Use case 3: accuracies in predicting the industry activity, obtained with the different distance metrics.

Methods	5	10	15	20	25	30	35	40	Mean
Ledger Account Overlap	0.358	0.310	0.358	0.307	0.313	0.324	0.347	0.341	0.332
Mahalanobis Similarity Score	0.232	0.276	0.207	0.295	0.197	0.239	0.227	0.232	0.238
Cosine Similarity Score	0.329	0.383	0.416	0.372	**0.390**	0.399	0.392	0.422	0.388
Tree Edit Distance	0.361	0.371	0.337	0.372	0.379	0.369	**0.406**	0.378	0.372
EMD-GraD	**0.378**	**0.410**	**0.427**	**0.440**	**0.390**	**0.411**	0.401	**0.425**	**0.410**

Note: The bold emphasis indicates the methods that achieve the highest accuracy for a specific vector dimension ranging from 5 to 40.

TABLE 9 Use case 4: accuracies in predicting the industry activity, obtained with the different distance metrics.

Methods	5	10	15	20	25	30	35	40	Mean
Ledger Account Overlap	0.381	0.424	0.443	0.405	0.432	0.430	0.427	0.417	0.420
Mahalanobis Similarity Score	0.192	0.154	0.185	0.169	0.155	0.179	0.175	0.180	0.174
Cosine Similarity Score	0.395	0.480	0.471	0.487	0.487	0.487	**0.491**	0.481	0.472
Tree Edit Distance	0.418	0.441	0.448	0.453	0.446	0.462	0.449	0.478	0.449
EMD-GraD	**0.429**	**0.492**	**0.522**	**0.512**	**0.506**	**0.498**	0.485	**0.511**	**0.494**

Note: The bold emphasis indicates the methods that achieve the highest accuracy for a specific vector dimension ranging from 5 to 40.

accuracy of 0.146. This is considerably lower than the accuracies from the stratified models in the other use cases because in this use case we have the largest number of industry groups and the smallest amount of training data. This suggests that using EMD-GraD as the basis for training the models provides a more effective solution in terms of accuracy, even when dealing with a large number of industry groups and limited training data.

4.4.7 | Conclusion

In this experiment, we propose a financial statement-based classification system for industry codes that utilizes a low-dimensional representation of a company as input. The representations are generated from the pairwise distances between companies using various financial statement distance metrics and are found to be transferable and easily adaptable for downstream tasks.

The results showed that EMD-GraD performed the best in all use cases based on the average industry classification accuracy. EMD-GraD performed on average 2.3% points better than the second-best metric across all use cases. The cosine method performed second-best in three out of four use cases, indicating that considering value information can outperform methods that only consider financial statement structure information. The tree edit distance metric performed second-best in one use case, suggesting that structure information may be more valuable in that particular case. The Mahalanobis similarity distance performed the worst in all use cases. In general, the best method performed 1.6 to 3.3 times better than the stratified model. Overall, the results indicate that financial statements provide valuable information for predicting industry sectors. Additionally, EMD-GraD was found to be effective when dealing with a larger number of industry groups and a limited amount of training data.

Our results confirm that EMD-GraD is superior to existing baselines as it takes into account both structural and value information. The lower-dimensional representation produced by EMD-GraD contains more useful information, as it outperforms methods that only consider a subset of this information. This supports our hypothesis that the inclusion of both structural and value information is beneficial for obtaining an effective low-dimensional company representation.

Furthermore, EMD-GraD is the most effective at capturing industry information in its low-dimensional representation. The company embedding obtained by using EMD-GraD is better able to distinguish companies based on their industry activity in a lower dimensional space. This is demonstrated by the fact that it is the most effective method for classifying the first two- and three-digit NACE codes for certain company subgroups.

5 | CONCLUSIONS AND FUTURE WORK

In this paper, we propose a new approach for selecting similar companies based on financial statements by introducing a new distance metric. Unlike current industry classification systems, EMD-GraD allows for a more flexible selection of peer companies by quantifying the similarity between two financial statements, taking into account both structure and value information. This is a novel approach within the field of peer company selection, as no existing methods combine both sources of information.

Our experimental analysis shows that EMD-GraD outperforms state-of-the-art approaches in identifying peer companies with high industry code overlap and low variance in financial ratios. This suggests that EMD-GraD is more effective at selecting peer companies that are financially related and engage in similar business activities. Additionally, EMD-GraD demonstrates superior performance in encoding companies into a low-dimensional space, capturing more industry information than existing methods. To our knowledge, this is the first company embedding based on financial statement information.

We also provide an efficient implementation of EMD-GraD, which significantly reduces the computational time and makes it more feasible to use the distance metric on a large scale. Furthermore, users are able to adapt the weight function w to their specific needs, allowing for a customizable version of EMD-GraD. Our results indicate that financial statements contain valuable information for predicting company activity and that our proposed company embedding leverages this information to improve the quality of downstream prediction tasks.

In future research, our work could be extended in several directions. One potential direction is to add an explainability layer that provides insight into how companies differ based on their traceable optimized flows, which could help better understand the underlying mechanisms that distinguish companies. Another potential extension is to represent companies as a time series of vertex-weighted trees, which would allow us to capture the evolution of a company over time. We also plan to explore the usefulness of EMD-GraD in other industries, such as bioinformatics, where data can be modeled as vertex-weighted trees. Additionally, since our company embedding is transferable and adaptable for downstream prediction tasks, we would like to investigate additional use cases where it could be beneficial, such as financial forecasting or risk assessment. Overall, this research has the potential to provide valuable insights into the performance of companies and could have significant implications for various industries.

ACKNOWLEDGMENTS

This research received funding from the Flemish Government, through Flanders Innovation & Entrepreneurship (VLAIO, project HBC.2020.2883) and from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program. This is an extended and revised version of a preliminary conference proceeding that was presented at the 2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr) (Noels et al., 2022). We would like to thank the Silverfin AI Team for their collaboration and support on this research project. Their expertise and insights were invaluable in helping us conduct our experiments and analyze our results.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Sander Noels https://orcid.org/0000-0001-6042-8461
Simon De Ridder https://orcid.org/0000-0003-0493-0160
Sébastien Viaene https://orcid.org/0000-0003-1001-8573
Tijl De Bie https://orcid.org/0000-0002-2692-7504

ENDNOTES

[1] https://www.silverfin.com

[2] https://www.silverfin.com

[3] https://github.com/snoels/earth-movers-graph-distance-metric

REFERENCES

Asche, F., & Misund, B. (2016). Who's a major? A novel approach to peer group selection: Empirical evidence from oil and gas companies. *Cogent Economics & Finance*, *4*(1), 1264538. https://doi.org/10.1080/23322039.2016.1264538

Berardi, G., Esuli, A., Fagni, T., & Sebastiani, F. (2015). Classifying websites by industry sector: A study in feature design. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, Association for Computing Machinery, pp. 1053–1059. https://doi.org/10.1145/2695664.2695722

Bernstein, A., Clearwater, S., & Provost, F. (2003). The relational vector-space model and industry classification. In *Proceedings of the Learning Statistical Models from Relational Data Workshop at the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 8–18.

Boyd, S., Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*: Cambridge University Press.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brown, S. V., Ma, G., & Tucker, J. W. (2021). Financial statement dissimilarity and sec scrutiny. Available at SSRN 3384394.

Bushman, R. M., & Smith, A. J. (2001). Financial accounting information and corporate governance. *Journal of Accounting and Economics*, *32*(1), 237–333. https://www.sciencedirect.com/science/article/pii/S0165410101000271

Cong, Y., Hao, J., & Zou, L. (2014). The impact of XBRL reporting on market efficiency. *Journal of Information Systems*, *28*(2), 181–207.

De Franco, G., Kothari, S. P., & Verci, R. S. (2011). The benefits of financial statement comparability. *Journal of Accounting Research*, *49*(4), 895–931. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-679X.2011.00415.x

Ding, K., Peng, X., & Wang, Y. (2019). A machine learning-based peer selection method with financial ratios. *Accounting Horizons*, *33*(3), 75–87. https://doi.org/10.2308/acch-52454

Fan, J. P. H., & Lang, L. H. P. (2000). The measurement of relatedness: An application to corporate diversification. *The Journal of Business*, *73*(4), 629–660.

Fang, F., Dutta, K., & Datta, A. (2013). LDA-based industry classification. In Baskerville, R. L., & Chau, M. (Eds.), *Proceedings of the International Conference on Information Systems, ICIS 2013*: Association for Information Systems, pp. 2500–2509. http://aisel.aisnet.org/icis2013/proceedings/ResearchInProgress/36

Gkotsis, P., Pugliese, E., & Vezzani, A. (2018). A technology-based classification of firms: Can we learn something looking beyond industry classifications? *Entropy*, *20*(11), 887.

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, *124*(5), 1423–1465. https://doi.org/10.1086/688176

Hoitash, R., Hoitash, U., Kurt, A. C., & Verdi, R. S. (2018). An input-based measure of financial statement comparability. Available at SSRN 3208928.

Hopkins, P. E. (1996). The effect of financial statement classification of hybrid financial instruments on financial analysts' stock price judgments. *Journal of Accounting research*, *34*, 33–50.

Jan, C. (2018). An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from taiwan. *Sustainability*, *10*(2), 513.

Jiang, S., Song, Z., Weinstein, O., & Zhang, H. (2020). Faster dynamic matrix inverse for faster LPS. https://arxiv.org/abs/2004.07470

Kahle, K. M., & Walkling, R. A. (1996). The impact of industry classifications on financial research. *Journal of financial and quantitative analysis*, *31*(3), 309–335.

Kanapickienė, R., & Grundienė, Z. (2015). The model of fraud detection in financial statements by means of financial ratios. *Procedia-Social and Behavioral Sciences*, *213*, 321–327.

Kee, T. (2019). Peer firm identification using word embeddings. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5536–5543.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Lee, C. M. C., Ma, P., & Wang, C. C. Y. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, *116*(2), 410–431. https://www.sciencedirect.com/science/article/pii/S0304405X15000197

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, *10*, 707–710.

Nagy, R. A., & Obenberger, R. W. (1994). Factors influencing individual investor behavior. *Financial Analysts Journal*, *50*(4), 63–68.

Noels, S., Vandermarliere, B., Bastiaensen, K., & De Bie, T. (2022). An earth mover's distance based graph distance metric for financial statements. In *2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFER)*, pp. 1–8.

Raman, N., Bang, G., & Nematzadeh, A. (2019). Multigraph attention network for analyzing company relations. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, Association for Computing Machinery, pp. 426–433. https://doi.org/10.1145/3373509.3373542

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a better measure of business proximity. *MIS Quarterly*, *40*(4), 1035–1056.

Vaidya, P. M. (1989). Speeding-up linear programming using fast matrix multiplication. In *30th Annual Symposium on Foundations of Computer Science*, pp. 332–337.

Vavasis, S. A. (1991). *Nonlinear optimization: Complexity issues*: Oxford University Press, Inc.

Yang, S., & Cogill, R. (2013). Balance sheet outlier detection using a graph similarity algorithm. In *2013 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFER)*, IEEE, pp. 135–142.

Yang, S. Y., Liu, F.-C., Zhu, X., & Yen, D. C. (2019). A graph mining approach to identify financial reporting patterns: An empirical examination of industry classifications. *Decision Sciences*, *50*(4), 847–876.

How to cite this article: Noels, S., De Ridder, S., Viaene, S., & De Bie, T. (2023). An efficient graph-based peer selection method for financial statements. *Intelligent Systems in Accounting, Finance and Management*, *30*(3), 120–136. https://doi.org/10.1002/isaf.1539