




A Graph Mining Approach to Identify Financial Reporting Patterns: An Empirical Examination of Industry Classifications

Steve Y. Yang [†] 
School of Business, Stevens Institute of Technology, 1 Castle Point on Hudson, Hoboken, NJ, 07030, e-mail: steve.yang@stevens.edu

Fang-Chun Liu 
Kate Tiedemann College of Business, University of South Florida-St. Petersburg, St. Petersburg, FL, 33701, e-mail: fangchun@mail.usf.edu

Xiaodi Zhu
Department of Finance at New Jersey City University, Jersey City, NJ, 07305, e-mail: xzhu@stevens.edu

David C. Yen
School of Economics and Business, State University of New York at Oneonta, Oneonta, NY, 13820, e-mail: David.Yen@oneonta.edu

ABSTRACT

This study proposes a quantitative method using the eXtensible Business Reporting Language financial accounting taxonomies to identify firms' common business characteristics and demonstrates that this graph mining approach can effectively identify industry boundaries. The premise of this method is based on the previous findings that financial accounts and the structural semantic information represented in financial statements reveal firms' general business operations and common characteristics if they have similar business models. Specifically, we introduce a graph similarity metric combined with spectral clustering algorithm to quantify the similarity of financial disclosures. Through industry classification comparison with the traditional classification schemes, the Standard Industrial Classification and the North American Industry Classification System, we show that the proposed method consistently clusters firms into their respective industries based on financial disclosures with significantly lower variance in a time-varying fashion. This novel graph mining method provides an automated way for decision makers to identify common business operations as well as detecting potential

The authors would like to thank the ICIS 2016 conference attendees for their valuable comments, and also thank Anzhela Knyazeva, Marco Enriquez, Roman Ivanchenko, Walter Hamscher, Mike Willias, Seung Won, and other 2016 internal seminar participants at the Division of Economic and Risk Analysis of the U.S. Securities and Exchange Commission for their valuable suggestions.

[Corrections added on November 13, 2018 after first publication on October 30, 2018: The second author Fang-Chun Liu's email address has been corrected.]

[†]Corresponding author.

financial fraud and uncovering accounting information misrepresentation.[Submitted: February 28, 2017. Revised: February 13, 2018. Accepted: September 25, 2018.]

Subject Areas: Financial reporting structure, Graph similarity, Industry classification, Semantic pattern, and XBRL.

INTRODUCTION

Financial statements are important sources of information showing the financial position and operational performance of a company (Bushman & Smith, 2001), and they are to fairly represent firms' business operations from a financial perspective. As a crucial factor influencing investors' decisions, the quality of financial statements conditioned on the extent of financial disclosures has long been the focus of academics, regulators, and those in the accounting profession (Schipper, 2007). The choice of different reporting structures, however, has raised the question of whether and how it affects information users' judgment about a firm's value. The lack of a reporting structural pattern detection mechanism has compounded the challenge of investigating the impact of reporting structural variances. To bridge this gap, the purpose of this study is to propose a graph mining approach to define industry boundaries, that is, industry classification, using financial statement reporting structures.

Throughout the last decade, the Securities and Exchange Commission (SEC) has invested much effort in improving the quality of financial reporting by mandating the adoption of the eXtensible Business Reporting Language (XBRL) (Janvrin, Pinsker, & Mascha, 2013; Perdana, Robb, & Rohde, 2014). XBRL facilitates better financial disclosure and easier information exchange that allow accounting information users to analyze a firm's performance and future growth opportunities in a timely manner (Hodge, Kennedy, & Maines, 2004; Liu, Luo, Sia, O'farrell, & Teo, 2014).ⁱ Standardization efforts in financial reporting have led to a large number of machine-interpretable vocabularies that can be employed to model the complex accounting practices in XBRL format. Despite increased transparency, accessibility, and richness of data after the XBRL mandate, it remains a challenge for many practitioners to automatically consume the semantic information for making better informed decisions, such as the ones in the financial services industry (Chowdhuri, Yoon, Redmond, & Etudo, 2014). To this end, this study utilizes XBRL to extract firms' financial statement structure information, and then investigates different presentation patterns of financial reporting choices to assess the effectiveness of identifying common business practices and operations.

ⁱ For instance, once the XBRL-encoded financial statement (i.e., "instance document") is filed with the SEC or uploaded to the company's Web site, investors are able to download the information in the financial statements within minutes and be in a position to start analyzing the data. The SEC viewer, located on the SEC's interactive data Webpage, is used to render the instance document (i.e., convert the XBRL-encoded information into human-readable form). Although the SEC viewer was made publicly available in late 2006, anecdotal evidence suggests that proprietary software to render instance documents was available prior to this time (SEC Release Nos 33-9002; 34-59324; 39-2461; IC-28609; File No.S&-11-08, see details at <https://www.sec.gov/rules/final/2009/33-9002.pdf>; accessed: 2016-09-01). Moreover, XBRL overcomes the interoperability issues traditionally associated with the information exchange across different platforms and software applications. By using XBRL, information can be exchanged seamlessly between these different systems (Boritz & No, 2005).

A common method for academic researchers and industry practitioners to control for industry effect is to rely on the industry classification to cluster firms with similar firm characteristics, such as business operations and products/services, into groups. The industry classification system therefore is important for information users in business as it provides a systematic way of identifying a group of similar companies (Kahle & Walkling, 1996; King & Slotegraaf, 2011; Narasimhan, Schoenherr, Jacobs, & Kim, 2015) and understanding the structure of economy (Christensen, 2013). It is well established that better industry classification contributes to better analysis results, compared with only considering firm size in the comparison (Kahle & Walkling, 1996). The most widely used industry classification scheme is the Standard Industrial Classification (SIC) codes, which was established in 1937 (Kolesnikoff, 1940). In 1997, the North American Industry Classification System (NAICS) was proposed and has gradually replaced the SIC codes (Pagell & Weaver, 1997). Meanwhile, several alternative industry classification systems have been developed and used by researchers and practitioners (Fama & French, 1997; Fan & Lang, 2000; Chan, Lakonishok, & Swaminathan, 2007; Chong & Zhu, 2012; Fang, Dutta, & Datta, 2013; Lee, Ma, & Wang, 2015; Hoberg & Phillips, 2016). The present study aims to extend this line of literature.

We propose a financial statement structure industry classification (FSSIC) method based on firms' financial statement similarity measured by the graph similarity metric. In this method, we combine the graph similarity method (Yang & Cogill, 2013) with the spectral clustering algorithm to define industry boundaries and validate the effectiveness of FSSIC through the comparisons with other established schemes, such as SIC and NAICS. To demonstrate the differences in firms' reporting structures, we first select four retail companies' balance sheets reported in XBRL format and show that the proposed method is able to reflect the unique firm characteristics of these four companies based on their choices of business models and resource allocation decisions. To further demonstrate the effectiveness of the proposed method, we conduct a large-scale data analysis by extracting corporate balance sheet data from the 10-K XBRL filings between 2010 and 2015 and clustering the constituent companies of the S&P 1500 into their corresponding industries. The results of comparisons between our proposed classification method and the two-digit SIC code classification show that the proposed FSSIC has lower intra-industry variance of selected financial ratios, that is, return on assets, return on equity, price-to-book ratio, and leverage ratio, which are all at statistical significance levels between 0.05 and 0.1. The results comparing with the NAICS classification show the same lower intra-industry variance on the same financial ratios. This overall indicates that the FSSIC can reliably classify firms with similar operating characteristics into their peer groups. Additionally, we compare FSSIC with another recent and comparable text-mining-based classification method, the fixed industry classification (FIC) method proposed by Hoberg and Phillips (2016). The comparison results show that FSSIC is complementary to FIC in which both have time-varying advantages over SIC and NAICS systems but with different emphases. To the best of our knowledge, this study is the first to explore the issue of varied financial reporting structures commonly observed in practice by applying the graph theory proposed in a related study (Yang & Cogill, 2013) to shed light on industry boundaries using companies' financial reporting structures.

Our study contributes to the existing literature in providing an alternative industry classification method that is based on firms' financial statement structures. This method is time-varying in nature. As firms change their business operations and strategies, their industry membership may change as a result of the resemblance with their peer group characteristics. Moreover, the graph mining approach presented in this study provides an automated means to identify common financial disclosure structural patterns and hence can be used as a potential method to help information users, including investors and regulators, to group firms based on reporting structures that inherently capture their common business operations and strategies. On the other hand, it can also help information users to identify distinct operational protocols, unique business strategies, or various fraudulent reporting practices. Given the overwhelming amount of data available to information users, we expect this proposed financial statement-based method to help users perform better data analysis and improve both the quality and efficiency of making business and investment decisions.

The rest of the article is structured as follows. The "Literature Review" section reviews the relevant literature discussing the impact of using different financial statement presentation techniques and the use of XBRL for better information sharing and disclosure quality, as well as literature on industry classification and alternative methods. The "Research Methodology and Data" section presents the similarity measure as a tree editing distance problem and then demonstrates how to combine it with the spectral clustering algorithm to solve the industry classification problem. In the "Research Result" section, we first present the research results and then discuss the findings supplemented with additional analyses and useful insights. Finally, the last section concludes the findings and elaborates on the contributions of this study.

LITERATURE REVIEW

To set the context of our research, this section first reviews the existing literature on the information representation of financial statements and its quality improvement using XBRL technology. We then review the prior work related to industry classification that leads to the motivation of proposing a financial statement-based industry classification method.

Financial Reporting and XBRL

Prior literature has found that the types of statement presentation styles influence human judgment (Stock & Watson, 1984; Vickery, Droge, & Markland, 1993). Financial statement structure involves many aspects of presentation such as classification, aggregation, placement, and labeling of financial items. Differences in these aspects of financial statement presentation affect investors' judgment (Frederickson, Hodge, & Pratt, 2006). For example, Hopkins (1996) shows that the balance sheet classification of financial instruments that include attributes of both debt and equity affects the stock price judgments of buy-side financial analysts. Similarly, Hirst and Hopkins (1998) show that reporting comprehensive income and its components facilitates detection of earnings management

by buy-side financial analysts and predictably affects their common stock price judgments. Maines and McDaniel (2000) document that alternative presentation of comprehensive income affects nonprofessional investors' evaluation on the disclosed comprehensive income information. Koonce, Lipe, and McAnally (2005) conduct a series of experiments and demonstrate that the financial statement items firms use to describe financial instruments and derivatives cause investors to assess economically equivalent instruments differently in terms of risks. Processing fluency theory suggests that with a more readable disclosure, investors may have a better perception of the reliability of disclosed information (Rennekamp, 2012). With the advancement of information technology, there is increased interest in utilizing technology to not only mitigate information asymmetry but also improve information processing efficiency (Cong, Hao, & Zou, 2014). Researchers have examined whether and how the adoption of different presentation formats, such as tabular format or portable document format (PDF), may affect users' perception of company performance (Benbasat & Dexter, 1986; Frownfelter-Lohrke, 1998; Lymer, 1999; Clements & Wolfe, 2000; Cong et al., 2014; Miller & Skinner, 2015).

One of the most notable reporting technological changes in recent decades is the adoption of XBRL (Baldwin & Brand, 2011). XBRL is an open-source standard that provides a mechanism to model business information and articulate the semantic meaning of reported financial concepts. As a markup language based on the eXtensible Markup Language (XML) and its syntax, XBRL can be used to define and facilitate the exchange of information contained in a company's financial statements by specifying a standardized framework to govern the definition of financial information (to humans and computers), its behavior, and associated characteristics. With XBRL, users can formulate customized calculations queries and to address specific questions that are of interest to users (Baldwin & Brand, 2011). XBRL advocates suggest that XBRL improves not only users' information search capability but also the information transparency (Hodge et al., 2004; Janvrin et al., 2013). For example, Yoon, Zo, and Ciganek (2011) find that the use of XBRL may improve the transparency and quality of business reporting by reducing information asymmetry. Accordingly, the use of XBRL is expected to help build a more stable and consistent reporting system that makes the use of financial information more efficient and effective (Pinsker & Li, 2008). More recently, Dhole, Lobo, Mishra, and Pal (2015) examine the implications of the SEC's XBRL mandate for financial statement comparability.

One topic that has not yet been investigated in previous research on financial reporting quality and XBRL, specifically at the intersection of improved comparability and XBRL usage, is the firm's choice of reporting structure that may reveal more insights on the firm's disclosure intentions as well as resource allocation decisions based on its business models and strategies (Holthausen & Leftwich, 1983; De Franco, Kothari, & Verdi, 2011). While firms with similar business operations should show great similarity in their financial statements, deviations, on the other hand, may reveal essential information about firms' unique business characteristics (Bradshaw et al., 2009; De Franco et al., 2011). In many cases, empirical evidence has shown that firms may use such financial reporting schemes to mislead information users for their own gains, which could result in manipulative earnings management and other forms of financial fraud (Hirst & Hopkins, 1998; Maines

& McDaniel, 2000; Frederickson et al., 2006). These behaviors will ultimately result in certain financial statement structural deviations from their peer groups. The patterns of financial reporting structures can then be examined by applying the graph theory because the nature of financial statements can be presented using tree structures. To fill the current research gap, we propose to apply graph mining methods on the structures of financial statement presentations, and use unsupervised machine learning techniques to cluster firms into respective industries, that is, industry classification.

Industry Classification

The purpose of industry classification is to divide firms into homogeneous markets with the assumption that firms in similar markets should exhibit similar firm characteristics, such as profitability and sales change (Clarke, 1989). However, the existing industry classification schemes, including the SIC and the NAICS, have two major limitations. One significant limitation is that industry classifications are not updated or modified very often although firms' business models have changed dramatically with rapid technological advancement and intense competition (Fan & Lang, 2000). Research indicates that current industry classifications cannot fully capture the change of the industry structure because of the rapid development of information technology and firm innovation (Segars & Grover, 1995; Carrillo, Druehl, & Hsuan, 2015). The infrequent updates of industry classifications keep the same industry scheme despite the market environment changes due to those new industries have emerged and/or old industries have diminished through technology and competition (Fang et al., 2013). The accuracy of existing industry classification schemes therefore has been questioned (Dalziel, 2007). Another limitation of existing industry classification schemes is the lack of identical and consistent classification methods. Researchers who extract data from different databases may find inconsistent results as different databases, such as COMPUSTAT and CRSP, have different defined industry codes for the same company (Kahle & Walkling, 1996).

Several studies have proposed different methods to classify companies into groups instead of solely relying on existing industry classification schemes. To measure the relatedness of companies, Fan and Lang (2000) use input–output (IO) tables and show that the results of IO tables outperform SIC codes. Lenard, Alam, and Booth (2000) classify firms based on their potential risks using fuzzy clustering algorithm. Chong and Zhu (2012) use tags in XBRL filings to group firms. Their findings show inconsistent grouping results with the NAICS classifications and suggest that the NAICS scheme is not as informative. Lee et al. (2015) propose to identify similar firms using Internet traffic patterns observed in the SEC EDGAR system (the official corporate filing platform managed by the SEC), which provides an alternative way of peer firm identification. Fang et al. (2013) apply a text mining technique, the Latent Dirichlet Allocation algorithm, to search firms' business description texts, and then cluster companies based on the relevance of business descriptions. A more recent text mining research addressing the issue of industry classification (Hoberg & Phillips, 2016) applies a clustering method to examine similarities of words disclosed in the business description section of annual reports

(10K). Hoberg and Phillips (2016) examine the intra-industry variation based on the text to show that their method of classifying industries is more informative than SIC and NAICS codes. To the best of our knowledge, an important attribute of examining the comparability issue, financial reporting structure, which presents firm's operational results and indirectly reflects the outputs of the chosen resource allocations and strategies, has not yet being examined in industry classification literature.

RESEARCH METHODOLOGY AND DATA

In this section, we first explain the graph similarity approach used for measuring financial statement structures, and then describe the industry classification method using the spectral clustering algorithm to define industry boundaries. We further define a methodology to validate the effectiveness of the proposed approach using balance sheet information extracted from corporate XBRL filing data distributed by the SEC.

Financial Statements Representation in a Tree Structure

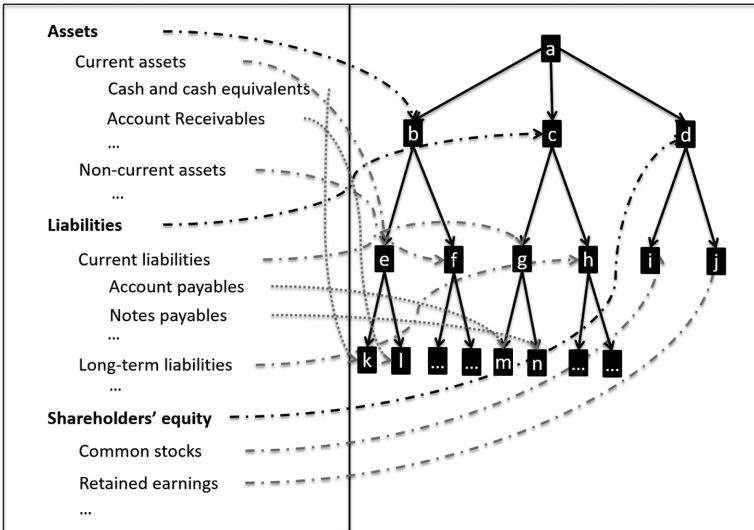
The presentation of accounting line items within a set of financial statements lends itself naturally to be represented by a labeled, directed, and rooted tree structure. In this study, we focus on the balance sheet for illustration purposes, as the same principle can be easily applied and extended to other financial statements. Generally speaking, a single financial statement or a set of financial statements can be modeled into one or multiple trees as illustrated in Figure 1. This formulation would provide a basis for constructing mathematical models and performing advanced data analysis on financial information structures.

Due to the lack of standardization, the automated use of financial statement information has been rather limited, and often it relies on time-consuming manual processing. To improve information processing efficiency, XBRL provides a standardized way to digitize financial statements so that the financial information of different companies, industries, and reporting periods can be normalized to perform an automated analysis. Using the XBRL technique, financial statements can be modeled in hierarchical structures along with additional semantic cross-associations and cross-references. Since the U.S. GAAP Taxonomy project initiated by the SEC in 2008, many U.S. companies started to publish their financial statements using the standard taxonomy and XBRL format. Under this new reporting requirement, each financial account or related financial concept is uniquely identified using either the standard U.S. GAAP Taxonomy code or the company's own specific code (if it is permissible under the U.S. GAAP guidance). In other words, vertices are labeled with unique labels in the tree formulation. Furthermore, the semantic relationships in the XBRL presentation among these financial concepts can be modeled in an ordered tree structure.

Graph Similarity Measure

The general graph similarity problem is NP-complete (Garey & Johnson, 1979). The proof is widely available (Garey & Johnson, 1979) and hence not provided

Figure 1: An example of balance sheet presented in a labeled, directed, and rooted tree graph.

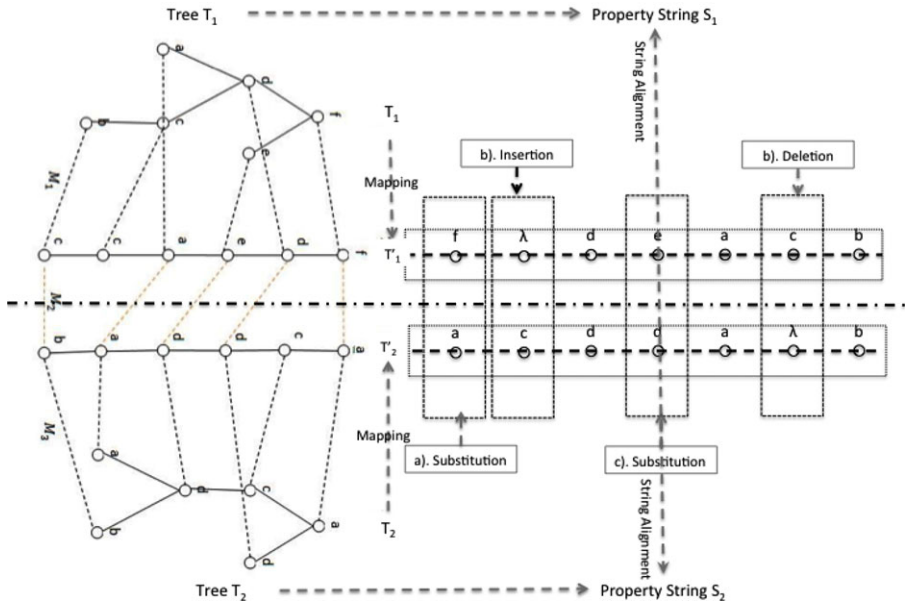


The left panel presents a typical balance sheet, and the right panel shows a hierarchical tree representing the balance sheet. The dashed arrows show the mapping mechanism to construct the statement into a tree structure. For example, “Assets,” “Liabilities,” and “Shareholders’ equity” are three items on the first level that are transferred into labels “b,” “c,” and “d” in the first children level under root “a” (balance sheet). The mapping then goes to the next level of balance sheet items until all items on the balance sheet are mapped.

in this article for the brevity purpose. Even though there is no polynomial time algorithm to solve this problem, there exist many approximation algorithms such as graph metrics. The standard algorithm for graph and subgraph isomorphism detection is the one introduced by Ullmann (1976). Since then, the maximum common subgraph detection has been addressed (Bunke & Allermann, 1983). Bille (2005) also conducted a survey about existing algorithms on the tree edit distance (Tai, 1979; Zhang & Shasha, 1989; Klein, 1998; Chen, 2001), alignment distance (Jiang, Wang, & Zhang, 1995; Jansson & Lingas, 2001; Yang & Cogill, 2013), and inclusion problems (Chen, 1998; Knuth, 1998) through the comparison of labeled trees based on simple local operations of deleting, inserting, and relabeling nodes.

Given that balance sheets can be represented as labeled, ordered trees, we are able to identify an appropriate metric for measuring the distance between pairs of labeled, ordered trees. We use the method proposed by Yang and Cogill (2013) to transform the underlying graphs into property strings and then align these strings with a dynamic programming algorithm (Levenshtein, 1966). By doing so, a tree edit distance problem can be reduced into a string edit distance problem, and an approximation to the tree edit distance can thus be obtained. Using the string edit distance, we obtain the similarity measure that can be computed polynomially in both time and space.

Figure 2: Alignment of two trees T_1 and T_2 .



This graph illustrates the graph similarity measure. The dashed horizontal line in the middle divides the process into the top transformation and the bottom transformation. The top transformation shows how the hierarchical tree T_1 is transformed into a property string S_1 , and the bottom shows the transformation from the tree T_2 to a property string S_2 . At the end, the tree edit distance problem between T_1 and T_2 represented on the left is solved as a string alignment problem between S_1 and S_2 represented on the right side of the graph.

It is noted that the graphs under consideration can be transformed into property strings in level order, where every node on a level is visited before going to a lower level. Here, we denote the root of the k th tree as $r_k^{T_k}$, and it has h_k levels and σ_{h_k} number of nodes at level h_k . The value of the property string at i th level and j th position for tree T_k is denoted as $v_{i,j}^{T_k}$. Hence, we can represent the property strings of tree T_1 and T_2 as follows (see Figure 2):

$$s_1 := r_1^{T_1} \circ v_{1,1}^{T_1} \circ v_{1,2}^{T_1} \circ \dots \circ v_{h_1,\sigma_{h_1}}^{T_1}, \quad (1)$$

$$s_2 := r_2^{T_2} \circ v_{1,1}^{T_2} \circ v_{1,2}^{T_2} \circ \dots \circ v_{h_2,\sigma_{h_2}}^{T_2}. \quad (2)$$

Figure 2 illustrates the concept of the graph similarity metric where the tree editing problem is cast into a string alignment problem. Following the definition and proof by Yang and Cogill (2013), we obtain Lemma 1, showing that the tree edit distance (δ_T) of original tree T_1 and T_2 has an upper bound that equals to the string editing edit distance (δ_S) of the transferred trees T'_1 and T'_2 . Moreover, this

upper bound can be obtained from the string alignment distance between S_1 and S_2 (T'_1 and T'_2) that can be computed using a dynamic programming algorithm, Levenshtein distance (Levenshtein, 1966). We focus on Levenshtein distance because of its importance among all the string matching algorithms (Navarro, 2001; Haque, Aravind, & Reddy, 2009). The binary sequence matching method based on Levenshtein (1966) has been adopted in dynamic programming algorithm for computing the string edit distance. It uses an $(n + 1) \times (m + 1)$ matrix, where n and m are the lengths of the two strings. The complexity of this algorithm is $O(nm)$. Other derivations of this algorithm can be readily incorporated for further enhancement.

Lemma 1: $\delta_T(T'_1, T'_2) = \delta_S(T'_1, T'_2)$, $\delta_T(T_1, T_2) \leq \delta_S(T'_1, T'_2)$.

The first part of the lemma, proven by Yang and Cogill (2013), tells us that the tree editing distance and string alignment between the two degenerated trees T'_1 and T'_2 are equivalent. It provides the basis for us to use the string alignment approach to solve the original tree edit problem. As illustrated in Figure 2, the tree edit distance between T'_1 and T'_2 on the left side and the string alignment between T'_1 and T'_2 on the right side in the graph near the center-dashed horizontal line are equivalent. The second part of the lemma shows that the string alignment cost $\delta_S(T'_1, T'_2)$ in general provides an upper bound for the tree edit distance cost $\delta_T(T_1, T_2)$ if the strings are the property strings of the corresponding trees whose similarity is to be measured. It provides us the basis to solve an NP-hard tree edit distance problem using an efficient string alignment algorithm. We henceforth construct an industry classification method based on this graph similarity measure.

Industry Classification Method

As discussed in the “Literature Review” section, prior research has proposed several industry classification methods to classify companies into groups. Based on the assumption that financial statement structures carry information about firms’ business characteristics and strategies, we develop a new time-varying FSSIC method using financial statements captured in XBRL format. We combine the graph similarity metric constructed earlier with the spectral clustering algorithm to group companies with similar reporting structural patterns into groups.

We first apply the graph similarity algorithm introduced in the “Graph Similarity Method” subsection on each pair of companies with the same fiscal year to produce an $N \times N$ similarity matrix among N companies. This similarity matrix $\mathbf{S}_{N \times N}$ represents pairwise similarity between N firms. Each element of this matrix \mathcal{S}_{ij} represents the similarity between firm i and firm j , and it follows symmetric property $\mathcal{S}_{ij} = \mathcal{S}_{ji}$. To reduce the computational complexity, we pick M benchmark firms. The similarity vector between firm i and the benchmark firms is used to represent the semantic structure feature of firm i . To pick the benchmark firms, we separate firms into groups using two-digit SIC codes, and only keep the firms with complete records in our research period from 2010 to 2015. We then choose one firm with the largest market capitalization from each two-digit SIC group, which produces 57 benchmark firms. The initial $N \times N$ similarity matrix is transferred into the $N \times M$ feature similarity matrix with $M = 57$.

We then apply the spectral clustering algorithm (Shi & Malik, 2000; Ng, Jordan, & Weiss, 2001) on the feature graph similarity matrix. Clustering is an efficient way to group data with similar features. Various clustering algorithms have been proposed and implemented by prior literature, including *k*-means clustering (Lloyd, 1982) and hierarchical clustering (Ward, 1963; Saeed, Malhotra, & Grover, 2011). Spectral clustering, a graph clustering method, is one of the most popular modern clustering algorithms (Shi & Malik, 2000; Ng et al., 2001). It is simple to implement and often outperforms traditional clustering algorithms such as the *k*-means algorithm. Previous studies that employ the spectral clustering algorithm are able to produce more accurate outputs (Chong & Zhu, 2012; Fang et al., 2013). Moreover, spectral clustering is a natural fit for our clustering problem, where the data points are clustered directly in the graph space using a graph similarity matrix.ⁱⁱ

Finally, we validate the effectiveness of our approach and compare the proposed industry classification with two-digit SIC codes (a total of 63 sectors) and three-digit NAICS codes (a total of 89 sectors), respectively. Because the number of clusters can potentially affect the cluster membership, we choose a number between the number of SIC and NAICS sectors as the input for the spectral clustering algorithm, which is $K = 75$. We also conduct robustness checks using different numbers of clusters $K = \{50, 55, 60, 65, 70, 75, 80\}$ to confirm that the results are robust across different cluster choices.

To conduct comparison between two industry classifications, we follow Guenther and Rosman’s (1994) study based on the assumption that intra-industry variance of financial ratios will be lower if an industry classification method can effectively group homogeneous companies into the same group. We choose several financial ratios (return on assets, return on equity, price-to-book ratio, and leverage ratio) commonly used by investors to make investment decisions (Bai, Hsu, & Krishnan, 2014) to validate our proposed financial statement–based industry classification. *F*-test is used to measure the difference of financial ratio variance, as shown in Equation (3). We then use the composite variance (*S*) proposed by Guenther and Rosman (1994) to construct the variance ratio for each industry classification scheme:

$$S = \frac{\sum_{i=1}^N (n_i - 1)V_i}{\sum_{i=1}^N (n_i - 1)}, \tag{3}$$

where *N* is the total number of industries, *n_i* is the total number of companies in industry *i*, and *V_i* is the variance of financial ratios of all companies in industry *i*.

ⁱⁱ The success of spectral clustering is mainly based on the fact that it does not make strong assumptions on the form of the clusters. As opposed to *k*-means, where the resulting clusters form convex sets (or, to be precise, lie in disjoint convex sets of the underlying space), spectral clustering can solve very general problems like intertwined spirals. Moreover, spectral clustering can be implemented efficiently even for large data sets, as long as the similarity graph is sparse. Once the similarity matrix is chosen, we then can solve a linear problem easily. There are no issues of dealing with local minima or restarting the algorithm for several times with different initializations. However, choosing a good similarity graph is not trivial, and spectral clustering can be quite unstable under different choices of the parameters for the neighborhood graphs.

The ratio of two composite variances from two different industrial schemes is used as the F -statistic:

$$F - \text{Statistic} = S_{FSSIC} / S_{benchmark}, \quad (4)$$

where S_{FSSIC} is the composite variance based on FSSIC, and $S_{benchmark}$ is the composite variance using one of the benchmark industry classifications, including two-digit SIC and three-digit NAICS and FIC from Hoberg and Phillips (2016). An F -statistic significantly less than 1 implies a lower intra-industry variance within FSSIC, which can lend support to the informativeness of the proposed method.

Data and Sample Description

We extract financial statement information in XBRL furnished under the U.S. GAAP Taxonomies, and then model balance sheets as a labeled, ordered tree. Each XBRL submission delivered to the SEC contains two distinct document sets, including an instance document and a taxonomy set. The instance document contains only the financial disclosure facts (alphanumeric, either numbers or letters). The taxonomy set is composed of a set of files, which defines financial accounting concepts along with the syntactic and semantic information about the financial concepts.ⁱⁱⁱ We utilize a commonly used open-source software, Arelle,^{iv} to process the XBRL submissions in order to extract the relevant data and information. Additionally, we use the application program interface (API) provided by Arelle to extract balance sheet presentation views for each set of XBRL submissions. As defined under the U.S. GAAP XBRL Taxonomy Architecture, each XBRL submission consists of four views of a financial disclosure, including presentation, definition, calculation, and label views. Each of these views is then represented by XBRL linkbases correspondingly.^v The software takes the semantic information captured in XBRL taxonomy files and provides an information set of the underlying relationships. Furthermore, combined with the information disclosed in the instance document, the software renders a tree representation of the semantic relationships for a particular submission instance.

We focus on S&P 1500 firms using the membership list of S&P 500 (large-cap), S&P 400 (mid-cap), and S&P 600 (small-cap) indexes at the end of December of the year prior to each fiscal year. We then collect annual reports (10K) of all S&P 1500 companies submitted to the SEC EDGAR system from 2010 to 2015, and use Arelle software to convert XBRL format submissions into accounting

ⁱⁱⁱ For more information about the U.S. GAAP XBRL Taxonomy, please refer to: <http://xbrl.us/sec-reporting/taxonomies/> (accessed: 2016-09-01).

^{iv} Arelle supports XBRL and its extension features in an extensible manner. It can be used as a desktop application and can be integrated with other applications and languages utilizing its Web service. Further information can be found at <http://arelle.org>.

^v Linkbase is defined using the XLink specification to represent a particular semantic relationship. For example, the presentation linkbase is to represent financial statement presentation relationships under the U.S. GAAP 2010 XBRL Taxonomy. For more information about the XBRL Linkbase specification, please refer to: <http://www.xbrl.org/Specification/xbrl-recommendation-2003-12-31+corrected-errata-2008-07-02.htm> (accessed: 2016-09-01).

Table 1: Data distribution summary.

| Fiscal Year | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|------|-------|-------|-------|-------|-------|
| Number of sample firms | 667 | 1,016 | 1,115 | 1,108 | 1,088 | 1,024 |
| Groups using two-digit SIC industry | 29 | 44 | 48 | 48 | 47 | 47 |
| Groups using three-digit NAICS industry | 41 | 56 | 59 | 58 | 59 | 58 |

Note: This table presents the data summary of sample firms from 2010 to 2015. The SEC XBRL filing rules were adopted in three phases: large companies began to submit their XBRL filings in 2009 and 2010, while all other companies were required to submit their XBRL filings for the period on or after June 15, 2011 (fiscal year 2011). As a result, the number of sample firms included in 2010 is much smaller than other years.

Table 2: Financial ratio definitions.

| Ratios | Definition |
|----------------------------|---|
| Return on Assets | Net operating income after depreciation (D178)/(Property, plant, and equipment (D8) + current assets (D4) – current liabilities (D5)) |
| Return on Equity | Net income before extraordinary items (D172)/Total common equity(D60) |
| Price-to-Book Ratio | (Close Price (D199) × Outstanding Shares (D61))/Total common equity (D60) |
| Leverage | Total liabilities (D181)/Total stockholders' equity (D216) |

Note: This table presents the definition of all financial ratios used in the following validation. All items used to calculate the financial ratios are extracted from COMPUSTAT database. The name of each item is presented in the *Definition* column followed by the item number in the parentheses.

concepts with a hierarchical structure. Furthermore, we collect both financial data and industry classification information from COMPUSTAT. Following Bhojraj, Lee, and Oler (2003), we drop firms with missing values on total assets, total long-term debt, net income before extraordinary items, debt in current liabilities, and operating income after depreciation. Moreover, to reduce the outlier effect, we only keep firms with a share price of more than \$3, net sales greater than \$100 million, and positive value on common stock and shareholders' equity. As a result, we obtain a total of 6,018 firm-year observations. Table 1 presents the distribution summary of the sample data.

To evaluate our industry classification method, we follow previous studies to validate our proposed FSSIC using the financial ratios that are commonly used to evaluate a firm's financial position and its value (Guenther & Rosman, 1994; Zopounidis, Doumpos, & Zanakis, 1999; Bhojraj et al., 2003). These ratios include return on assets, return on equity, the price-to-book ratio, and the leverage ratio. The variable definitions and the data sources in the COMPUSTAT database are summarized in Table 2. Table 3 provides the descriptive statistics of the selected financial ratios.

Table 3: Descriptive statistics of the selected financial ratios.

| Ratios | Obs. | Mean | Std. | 25th Percentile | Median | 75th Percentile |
|----------------------------|-------|-------|--------|-----------------|--------|-----------------|
| Return on Assets | 5,873 | 0.274 | 4.582 | 0.112 | 0.215 | 0.366 |
| Return on Equity | 6,018 | 0.159 | 1.780 | 0.068 | 0.122 | 0.192 |
| Price-to-Book Ratio | 6,018 | 4.687 | 27.066 | 1.624 | 2.461 | 3.924 |
| Leverage | 6,018 | 2.986 | 37.684 | 0.619 | 1.109 | 1.985 |

Note: This table presents the summary statistics of the defined variables downloaded from COMPUSTAT database.

RESEARCH RESULTS

This section first illustrates the financial statement pattern identification using the graph similarity metric by analyzing four retail firms' asset sections within their balance sheets. We then apply the method proposed in the "Industry Classification Method" subsection on the S&P 1500 companies and examine the effectiveness of the proposed method with the two traditional industry classification schemes, SIC and NAICS. Finally, we use the retail industry to examine the changing nature of the time-varying classification method.

Pattern Detection Illustration

To demonstrate the effectiveness of the algorithm, we selected four retail companies (Costco, Macy's, Target, and Walmart) with distinct business features and then applied the similarity measure to test whether our approach effectively captures the major differences of their corresponding balance sheet structures. Companies competing in the same industry may implement different business practices or strategies to improve their short-term business performance and sustain their competitive advantage in the long run. We expect that such business differences will be reflected in their financial statement structures. As a result, pair comparisons of financial statements across firms are expected to reveal the unique patterns of reporting structural differences, which indirectly provide insights in firms' business strategies.

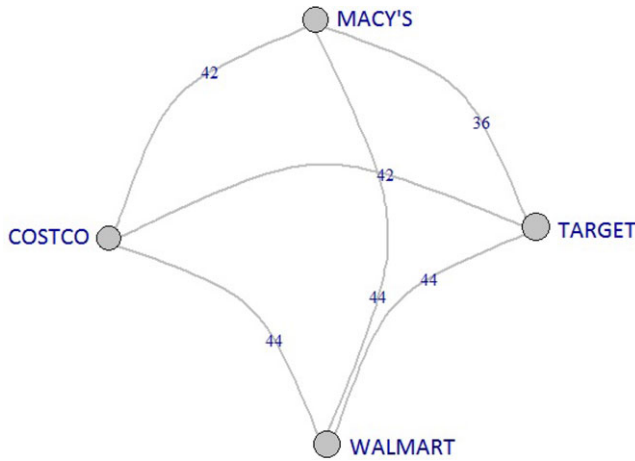
Figure 3 presents the comparison of the semantic structures extracted from the four companies' asset sections of their balance sheets filed in XBRL format. To highlight the structural differences across firms, we align the concepts representing similar accounting items in the same row. Each company has two hierarchical levels of accounting concepts in the asset section. The concepts in level 1 perform as a summary item for level 2 concepts. To visually differentiate different levels, we distinguish concepts in level 1 using red font and level 2 using black font. Most of the accounting concepts start with "us-gaap," which is a prefix for U.S. GAAP taxonomy that is the standard taxonomy used by the SEC for XBRL submission. Comparing these semantic structures, all four companies share four common items, including "us-gaap:CurrentAssets," "us-gaap:Property-PlantAndEquipmentNet," "us-gaap:OtherAssetsNoncurrent," and "us-gaap:Assets." Some items are associated with a specific business strategy of one company that is not common to the others. For example, Costco has a unique item "us-gaap:ShortTermInvestments"

Figure 3: This figure presents the asset section of four retail firms’ balance sheets in 2014: Costco, Macy’s, Target, and Walmart.

| Costco | | Macy’s | | Target | | Walmart | |
|---|---------|---|---------|---|---------|---|---------|
| level 1 | level 2 | level 1 | level 2 | level 1 | level 2 | level 1 | level 2 |
| us-gaap:AssetsCurrentAbstract | | us-gaap:AssetsCurrentAbstract | | | | us-gaap:AssetsCurrentAbstract | |
| us-gaap:CashAndCashEquivalentsAtCarryingValue | | us-gaap:CashAndCashEquivalentsAtCarryingValue | | us-gaap:CashCashEquivalentsAndShortTermInvestments | | us-gaap:CashAndCashEquivalentsAtCarryingValue | |
| us-gaap:ShortTermInvestments | | | | | | | |
| us-gaap:ReceivablesNetCurrent | | us-gaap:ReceivablesNetCurrent | | | | us-gaap:ReceivablesNetCurrent | |
| us-gaap:InventoryNet | | us-gaap:InventoryFinishedGoods | | us-gaap:InventoryNet | | us-gaap:InventoryNet | |
| us-gaap:DeferredIncomeTaxesAndOtherAssetsCurrent | | | | | | | |
| | | us-gaap:PrepaidExpenseCurrent | | | | us-gaap:PrepaidExpenseAndOtherAssetsCurrent | |
| | | | | us-gaap:AssetsOfDisposalGroupIncludingDiscontinuedOperationCurrent | | us-gaap:AssetsOfDisposalGroupIncludingDiscontinuedOperationCurrent | |
| | | | | us-gaap:OtherAssetsCurrent | | | |
| us-gaap:AssetsCurrent | | us-gaap:AssetsCurrent | | us-gaap:AssetsCurrent | | us-gaap:AssetsCurrent | |
| us-gaap:PropertyPlantAndEquipmentNetAbstract | | | | us-gaap:PropertyPlantAndEquipmentNetAbstract | | us-gaap:PropertyPlantAndEquipmentNetAbstract | |
| us-gaap:Land | | | | us-gaap:Land | | | |
| us-gaap:BuildingsAndImprovementsGross | | | | us-gaap:BuildingsAndImprovementsGross | | | |
| us-gaap:FixturesAndEquipmentGross | | | | us-gaap:FurnitureAndFixturesGross | | | |
| | | | | tgt:CapitalizedComputerHardwareandSoftwareGross | | | |
| us-gaap:ConstructionInProgressGross | | | | us-gaap:ConstructionInProgressGross | | | |
| us-gaap:PropertyPlantAndEquipmentGross | | | | | | us-gaap:PropertyPlantAndEquipmentGross | |
| us-gaap:AccumulatedDepreciationDepreciationAndAmortizationPropertyPlantAndEquipment | | | | us-gaap:AccumulatedDepreciationDepreciationAndAmortizationPropertyPlantAndEquipment | | us-gaap:AccumulatedDepreciationDepreciationAndAmortizationPropertyPlantAndEquipment | |
| us-gaap:PropertyPlantAndEquipmentNet | | us-gaap:PropertyPlantAndEquipmentNet | | us-gaap:PropertyPlantAndEquipmentNet | | us-gaap:PropertyPlantAndEquipmentNet | |
| | | | | | | wmt:PropertyUnderCapitalLeaseNetAbstract | |
| | | | | | | us-gaap:CapitalLeasedAssetsGross | |
| | | | | | | us-gaap:CapitalLeasesLesseeBalanceSheetAssetsByMajorClassAccumulatedDepreciation | |
| | | | | | | us-gaap:CapitalLeasesBalanceSheetAssetsByMajorClassNet | |
| | | | | us-gaap:DisposalGroupIncludingDiscontinuedOperationAssetsNoncurrent | | | |
| | | us-gaap:Goodwill | | | | us-gaap:Goodwill | |
| | | us-gaap:IntangibleAssetsNetExcludingGoodwill | | | | | |
| us-gaap:OtherAssetsNoncurrent | | us-gaap:OtherAssetsNoncurrent | | us-gaap:OtherAssetsNoncurrent | | us-gaap:OtherAssetsNoncurrent | |
| us-gaap:Assets | | us-gaap:Assets | | us-gaap:Assets | | us-gaap:Assets | |

The accounting concepts are presented in a hierarchical structure where parent concepts are placed on level 1 and detailed concepts are placed as children (level 2) of the corresponding parent item. We use red and black colors to distinguish accounting concepts at level 1 and level 2, respectively. Most of the accounting concepts are from standard U.S. GAAP taxonomy with prefix “us-gaap,” and only two concepts are customized with a unique prefix related to the specific firm “tgt” (Target) and “wmt” (Walmart). We use shade gray to highlight firms’ customized concepts. The accounting elements ending with “abstract” are only used for grouping purposes.

Figure 4: Comparative analysis of four retail companies' 2014 balance sheets: Costco, Macy's, Target, and Walmart.



The circles represent the balance sheet structures of the selected firms, and the lines between them represent the similarity between the two linked firms. The numbers tagged on the lines represent the exact similarity scores.

which the other three companies do not have. A close examination of the details of accounting items can also provide additional information about a company's business. For example, under the property, plant, and equipment section, all companies provide detailed items of their long-term assets except for Macy's. Such a difference, through the examination of the fixed assets that are included in this property, plant, and equipment section, captures and signifies that Macy's capital structure and the associated capital expenditure strategy are different from the other three. "tgt:CapitalizedComputerHardwareandSoftwareGross" and "wmt:PropertyUnderCapitalLeaseNetAbstract" are customized tags defined by Target and Walmart. The prefix of these two customized tags, that is, "tgt" and "wmt," are firm-specific concepts. The use of firm-specific concepts indicates that this firm has certain unique business characteristics that cannot be captured by a standard tag provided by the SEC. The identification of firm-specific tags can potentially signal unique business attributes of a firm to help information users better understand a firm's practice.

We then create an adjacency matrix among the four balance sheets of the selected retail companies in Figure 4. An edge between any two vertices represents the similarity between the two companies. In addition, the labels on the edges are the distance measures of these balance sheets. These companies carry many similar items with similar structures. However, there are several major distinct items among themselves that are consistent with the observed companies' strategies:

- (1) Macy's has a very simple "us-gaap:PropertyPlantAndEquipmentNet" section compared with the other three companies.

- (2) Walmart has an elaborate section for capital leases under the item entitled “wmt:PropertyUnderCapitalLeaseNetAbstract,” which contributes to the largest distance between itself and the other three companies. This finding is consistent with Walmart’s extensive use of leasing properties to offer physical store coverage.
- (3) Target distinguishes itself with a high-level asset item called “us-gaap:LoansHeldForSaleConsumerCreditCard.” This is aligned with its strategy in offering firm-issued credit cards as an alternate source of revenue.
- (4) Costco uniquely separates itself from others with the membership related items such as “us-gaap:CustomerRefundLiabilityCurrent” and “us-gaap:OtherDeferredCreditsCurrent” under the “Current liabilities” section. The membership program is a focal strategy that Costco adopts to enhance its brand awareness and customer loyalty.

Another important observation is that many of the large selection of the U.S. GAAP concepts can be used interchangeably and even by doing so, the differences turn out to be very subtle. This is the reason why we observe the distances among these four retail companies are actually larger than the unit distance of the basic tree editing operations. In order to further enhance the performance of this approach, one may consider adding a preprocessing step with which similar concepts can be unified as a single concept. In this way, this practice will reduce the noise generated by the nuance of the similar concepts. Taken together, we can clearly identify the distinct business features and patterns being highlighted by this similarity measure in general.

Industry Classification and Validation

In this section, we provide validation results using *F*-test by comparing our proposed FSSIC results with the existing industry classification schemes. We calculate the composite variance *S* on each industry-year observation after removing industries containing less than five companies. Then, we apply *F*-test on the selected financial ratios defined in Table 2. The results of our *F*-test are shown in Table 4. Figures 5 and 6 present a visualization of the clustering results for all the firms in 2015. From these figures, we are able to observe that the clusters are distinctively different, and the sizes of these groups are relatively consistent. One can also identify the memberships of the benchmark companies in different clusters through Figure 6.

The comparison result between the FSSIC and the two-digit SIC industry classification shows that our proposed method has lower intra-industry variance across ratios, and these results are statistically significant at the 0.05 and 0.10 levels, respectively. Similarly, the comparison results with the three-digit NAICS industry classification support the effectiveness of our proposed method. These results overall indicate that the proposed financial statement structure-based industry classification method is able to effectively classify firms with more similar operating characteristics into the same group. In other words, our proposed method can be used as a better industry classification approach than the commonly used

Table 4: Comparison of variances between industry classification schemes.

| Ratios | F-Statistic | Degree of Freedom |
|------------------------|-------------|-------------------|
| FSSIC vs. SIC | | |
| Return on Assets | 0.960* | 5,395 and 5,407 |
| Return on Equity | 0.948** | 5,515 and 5,550 |
| Price-to-Book Ratio | 0.962* | 5,515 and 5,550 |
| Leverage | 0.966* | 5,515 and 5,550 |
| FSSIC vs. NAICS | | |
| Return on Assets | 0.787*** | 5,126 and 5,407 |
| Return on Equity | 0.955** | 5,261 and 5,550 |
| Price-to-Book Ratio | 0.869*** | 5,261 and 5,550 |
| Leverage | 0.922*** | 5,261 and 5,550 |
| FSSIC vs. FIC | | |
| Return on Assets | 0.942** | 5,385 and 5,211 |
| Return on Equity | 1.013 | 5,528 and 5,333 |
| Price-to-Book Ratio | 0.961* | 5,528 and 5,333 |
| Leverage | 1.486 | 5,528 and 5,333 |

Note: The result above is obtained by winsorizing data at 1% and 99%, respectively, for ruling out the outlier effects.

*** indicates significant level at 1%.

** indicates significant level at 5%.

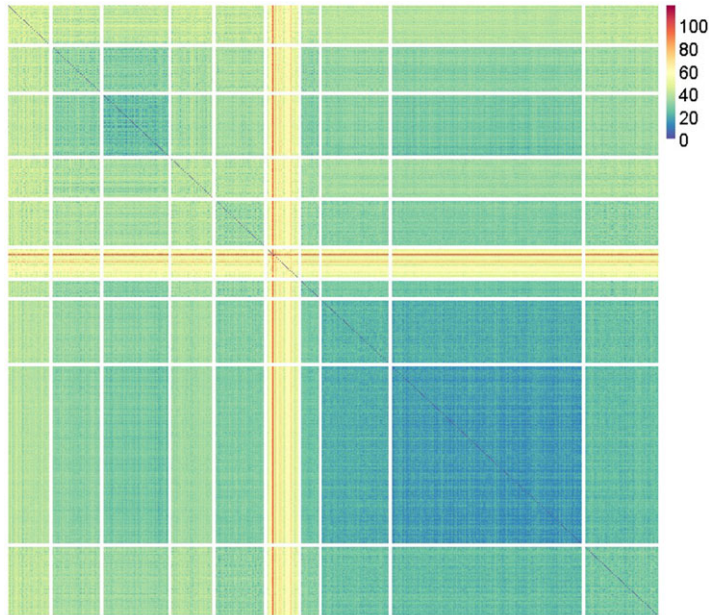
* indicates significant level at 10%.

industry grouping methods, such as the two-digit SIC or three-digit NAICS codes. Aside from complementing our understanding of firm characteristics through the qualitative attribute of reporting structure information, we expect the proposed method, which has been proven to provide a more accurate scheme on grouping firms, potentially benefits research that addresses questions such as firm comparisons and industrial effect controls.

Hoberg and Phillips (2016) propose two classification schemes, the Text-based Network Industry Classifications (TNIC) and the FIC, where TNIC is more informative than FIC. However, TNIC is firm-centric classification that requires a central firm to build the group of peer firms, and it loses the transitivity property that two similar firms may have different groups of peer firms. Due to these limitations of TNIC structure, it is not a comparable classification with our proposed FSSIC. Thus, we only perform comparison with FIC. To match the number of industries, we choose FIC with 100 industries for the comparison.^{vi} The last panel in Table 4 shows the *F*-test results between the two classifications. FSSIC has lower intra-industry variance than FIC on Return on Assets and Price-to-Book ratio. It is clear that these two methods are comparable to each other in terms of classification performances. Each has its own unique emphasis. However, we argue that FIC suffers two limitations compared with FSSIC. First, FIC is based on business descriptions that only consider firms' products and services. It can be very dynamic

^{vi} We download the FIC data from Hoberg-Phillips Data Library <http://hobergphillips.usc.edu/industryclass.htm>. In this analysis we use FIC100.

Figure 5: This figure presents a heatmap based on pairwise dissimilarity among all firms in FY2015.



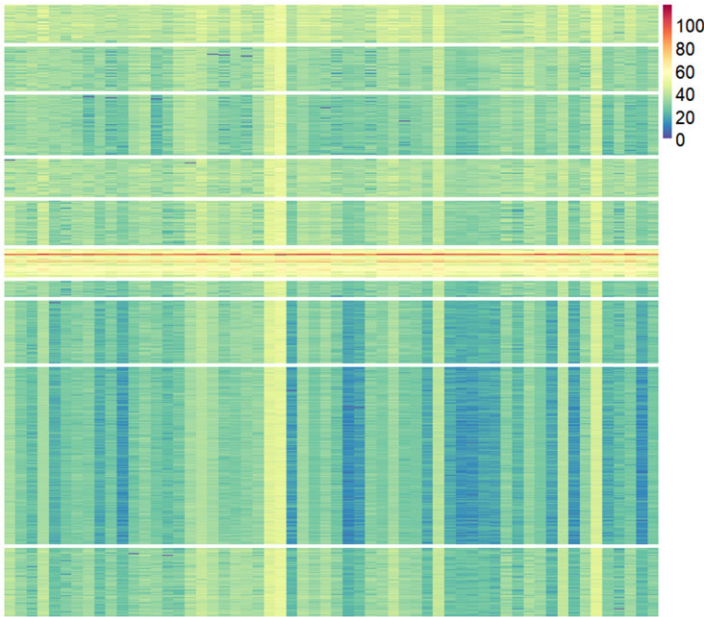
The heatmap shows that companies that have similar pairwise dissimilarity are grouped into one cluster. There are a total of 10 distinct clusters (the squares along the diagonal line with different colors). The sizes of the squares along the diagonal line indicate the sizes of the clusters, and the color of the clusters indicates the similarity across different clusters.

in nature that may result in significant inconsistencies over time. For example, innovative products may appear very different, but they should be relatable to the existing market with similar competitors over time. The text mining-based FIC will not be able to capture such semantic linkages. Therefore, classification based solely on the descriptions of the products and services can be misleading sometimes. Second, FIC is based on the descriptive data without a standard format. The unstructured texts require more effort on data extraction and processing, while FSSIC is based on financial statements that are well structured in XBRL format with standard taxonomies. Therefore, we argue that FSSIC and FIC methods can be complementary to each other in identifying industry peer groups.

In addition, we provide an example to illustrate the time-varying nature of the proposed FSSIC comparing with the other three classifications. We apply the spectral clustering algorithm with $K = 10$ on the feature semantic similarity matrix to compare with one-digit SIC industries, two-digit NAICS industries, and FIC industries with 25 clusters.^{vii} To match FSSIC and FIC clusters to a specific

^{vii} We download the FIC data from Hoberg–Phillips Data Library <http://hobergphillips.usc.edu/industryclass.htm>. The data with the smallest number of clusters are FIC25 that has 25 clusters. In this analysis, we use FIC25.

Figure 6: This figure presents a heatmap based on the dissimilarity between firms in FY2015 and the benchmark companies.

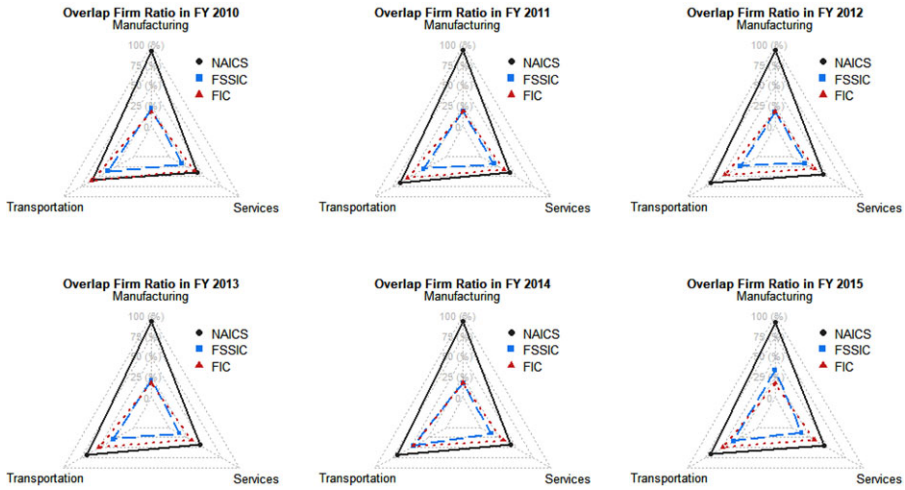


The 10 clusters including all firms in FY2015 are shown along the vertical axis. The benchmark companies are shown along the horizontal axis. The heatmap indicates that companies that have similar relationships with all benchmark companies are labeled in one cluster. The color of the intersection of the cluster and the benchmark indicates the similarity between them.

industry, we pick a cluster i with the largest number of overlapped firms as the closest cluster to a one-digit SIC industry j . This example uses the retail sector with SIC codes between 5200 and 5999. For each year, we pick one FSSIC cluster and one FIC cluster following the matching rule we define earlier. We also include the retail industry based on two-digit NAICS code. We observe that companies in the retail sector are clustered into different FSSIC clusters from 2010 to 2015. Figure 7 presents a visualization of the overlap ratio of all four industry classification systems, that is, SIC, NAICS, FIC, and FSSIC. In this visualization, we use SIC as the baseline and measure the overlap on the three largest industries, that is, manufacturing, transportation, and services. The larger the corresponding triangle area covers, the bigger the overlap between the specified methods with SIC becomes. From the graph, we see that NAICS has the largest overlap with the baseline SIC classification, while FSSIC and FIC have relatively similar overlap ratios against SIC classification.

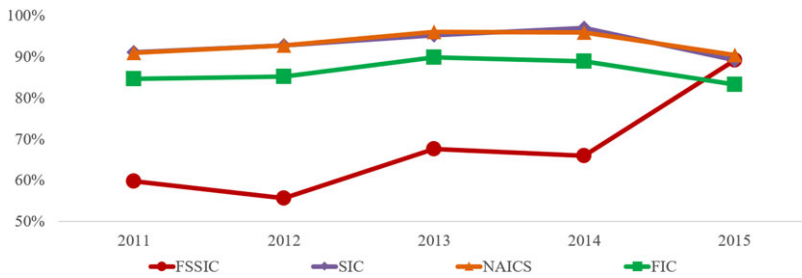
Next, we compare the change of membership over time for the four classification methods. Figure 8 provides a better illustration of the overall industry classification results across different years using preserved ratio, which is the ratio of companies in year $t - 1$ preserved to year t to the total number of firms in year

Figure 7: The figure presents the overlap firm ratio over 6 years for three major one-digit SIC industries with the largest number of firm population, including manufacturing, transportation, and services.



The overlap firm ratio is calculated as $N^{i}_{overlap}/N^{i}_{SIC} \cdot /N^{i}_{SIC}$ is the total number of firms in one-digit SIC industry i . $N^{i}_{overlap}$ is the number of overlapped firms between the SIC industry and cluster industry using one of the corresponding industry classifications, that is, NAICS, FSSIC, or FIC. The percentage value shown in the figure represents the value of overlap firm ratio with the maximum value 100%.

Figure 8: The figure shows the change of firms in retail sector using one-digit SIC, two-digit NAICS, FIC, and FSSIC.



The y-axis presents the value of preserved ratio at year t , which is the ratio calculated as N^{pre}_t/N^{all}_{t-1} , where N^{pre}_t is the number of companies preserved in $Year_t$ from $Year_{t-1}$, and N^{all}_{t-1} is the total number of companies in $Year_{t-1}$.

$t - 1$. The result shows that SIC, NAICS, and FIC have similar preserved ratio, above 80%, for the whole period, while FSSIC has lower preserved ratio than the other three classifications. From 2010 to 2014, 60–70% of firms are preserved in the same FSSIC retail cluster. The result indicates the consistency of membership in the retail cluster. In other words, most firms adopt similar balance sheet

semantic structures over time. Compared with SIC and NAICS, FIC has lower preserved ratio indicating the time-varying feature of FIC. However, the difference is small, suggesting that FIC retail cluster does not have significant change during the whole time period. In summary, the results demonstrate the time-varying feature of FSSIC. The dynamic classification preserves the companies with similar balance sheet semantic structures in one group, and ensures a highly homogeneous group by removing dissimilar companies and adding new ones.

Discussion

In the “Industry Classification and Validation” subsection, we demonstrate that the proposed graph mining method is an effective approach to identify balance sheet structural patterns with a high level of accuracy. Application of such an approach to a large amount of financial statements can become an effective way of helping accounting information users to identify their own investment interests toward the analyzed firms. A high-quality financial report should genuinely reflect the operational nature of the business, the competition, and the sensitive fluctuation of the industry environment. If we assume that financial statements reflect the “truth” of the underlying businesses, the proposed FSSIC method is a natural way to define the industry boundaries in a time-varying fashion. As economic conditions change, firms may adjust their business operations and hence deviate from their previous practices. Such deviation could help signal valuable investment opportunities or hidden information to investors and other decision makers. The representation of business strategies should be reflected through the examination and appreciation of similarities and differences of financial reporting structures. Investigating financial reporting structure differences over time will provide a possible way for decision makers to effectively and efficiently identify the changes of a firm’s operations and potentially pinpoint valuable investment opportunities associated with the observed changes. If certain patterns of business structures are correlated with positive business returns, investors would be able to use this proposed method to identify these firms and consequently make better investment portfolio choices accordingly.

As shown in the “Industry Classification and Validation” subsection, the proposed FSSIC method may be an alternative method to group homogeneous companies effectively. Compared with existing SIC and NAICS classification schemes, the proposed method is able to track the change of a company’s business over time and cluster the company into a more homogeneous group based on the commonalities detected through reporting structures. It overcomes the major drawback of the current classification schemes. Moreover, previous literature also proposed different industry classification methods that are more informative than existing schemes (Fang et al., 2013; Hoberg & Phillips, 2016). In the comparison with FIC classification proposed by Hoberg and Phillips (2016), FSSIC has lower intra-industry variance of return on asset and price-to-book ratios. However, FIC relies heavily on the complex qualitative information contained in the business description section that is, to some extent, difficult to implement in practice. Specifically, FIC uses business description data that concentrate on the products and services, while the proposed FSSIC method considers all underlying

business activities represented by the structure of financial statement. FSSIC thus uses more comprehensive information to measure the similarity between firms. As explained and demonstrated in the article, the proposed graph mining method only uses the semantic structure of financial statements that can be easily accessed by information users using XBRL. Semantic structure not only measures the similarity between two companies, but also provides information related to a company's business strategy that is often not specified in the business description section.

The proposed methodology may also provide opportunities for regulators to identify the conflict of potential reporting issues. For example, new business practices could be identified when compared with its own history and its industry peers using the deviation measure. The automated comparison analysis proposed in this study could also be used to help identify XBRL technical errors for better XBRL improvements. The subsequent corrections of these errors are expected to improve financial reporting quality. If the reporting choices are manipulated for some reason, any deviation from the "truth" may actually signal potential fraud or unintended mistakes. The identification of these deviations based on the industry benchmark is thus expected to provide an effective way to improve financial information quality and detect potential fraud.

However, we also recognize certain limitations of this study. First of all, we take note that there is a learning curve for firms to adopt the XBRL technology in furnishing their financial statements. The mandatory program started in 2009 and it was completed in 2011. We therefore have smaller samples in 2010 and 2011 than other periods. Over the course, we observe the quality of the XBRL reports varied due to factors such as vendor software change, taxonomy updates, use of custom tags, etc. These factors inevitably introduce errors to the structure similarity measures and hence affect the quality of industry classification. We foresee that these issues will gradually go away, and the similarity measure will be truly reflective of firms' real financial operations and business characteristics as firms gain more experience in filing XBRL reports in the future. Second, we only examine the structural patterns of the balance sheet in identifying common industry characteristics. It is conceivable that other financial statements such as income statement, shareholder's equity statement, and footnotes will also provide useful information for differentiating companies' business operations and strategies. To further enhance the proposed method, we suggest future studies to consider a more comprehensive graphical model where forests of trees will be considered in measuring the similarity matrix. Furthermore, future research could also consider the materialization of financial concepts in capturing significant business activities. In the current proposed method, we treat all the financial concepts equally. But in reality, there are times firms may report certain business activities with insignificant values compared with other activities. Removing less important financial concepts will inevitably remove noise activities and further enhance the performance of the industry classification results. The purpose of the present article is to propose a financial statement structure-based industry classification method with combination of a graph similarity algorithm and a spectral clustering method. We aim to define a clear scope for this study and make its contributions to the existing literature clear. Therefore, we document these improvements for future studies.

CONCLUSION

The study presents a graph mining method, specifically a graph similarity clustering method, to identify commonality in firms' financial statement structures. In the case of XBRL representation of financial disclosures, the method can effectively identify industries where companies opt to use a certain disclosure structure to represent their business operations and strategies. We study the S&P 1500 companies' balance sheet structures over a 6-year period, from 2010 to 2015, and show that commonality can be identified through clustering firms' individual choices of reporting structures. Furthermore, we demonstrate that when used for defining industry groups, the proposed financial statement-based industry classification method outperforms the existing SIC and NAICS industry classification methods in identifying firms' business homogeneity. Overall, we show that the proposed graph similarity clustering method applied to financial statement structures is sensitive in identifying firms' business operations and strategies. This study presents a novel, automated way of measuring accounting information presented in the financial disclosures. We believe that the method will benefit and advance research, in particular in the areas of financial reporting quality and XBRL, for conducting large-scale data analyses of financial statement information.

Finally, our results using actual corporate XBRL filing data suggest that the industry classification can be refined by considering firms' reporting choices that affect financial statement structures. Firms' reporting choices could be motivated by a variety of reasons such as economic conditions, strategic considerations, earnings management, or managerial styles, etc. Future research may employ alternative research methods, such as field study or survey, to further investigate firms' decisions in using different financial reporting structures. For example, we posit that under the similar economic conditions, firms in similar business areas would adopt similar accounting structures and use similar financial items in their financial statements. Surveying managers across different firms or conducting a field study to understand how firms make their accounting choices based on their business decisions under different economic conditions would further validate the financial statement-based industry classification approach.

Based on our discussions with various financial accounting standard setters, financial regulators, and accounting auditors, this proposed method helps practitioners discover reporting structural patterns embedded in a large amount of complex financial disclosures. For regulators such as the Company House/HMRC in the United Kingdom, the Australia Taxation Office, and the National Tax Agency in Japan who collect the massive amount of XBRL disclosures for taxation and regulatory purposes, this approach can be extremely useful to help them to mine unusual disclosures such as potential tax evasions and accounting frauds. For accounting standard setters such as the Financial Accounting Standards Board (FASB), this technique will enable them to automatically identify the specific use of custom XBRL elements by industry groups, as we demonstrated in the experiment where extraordinary deviations from the industry norms are normally caused by heavy use of custom tags. Therefore, XBRL taxonomy designers can identify the commonly used custom elements and generate new standard elements in the next version of the taxonomy. By doing so, we believe that data interoperability

will be improved over time. For auditors, the proposed approach, which examines the qualitative aspect of financial reporting quality, can complement current audit analytic practice to help identify significant changes of accounting items. If an audit client's financial statement structure is found to be significantly different from its industry common practices, this variation can signal the potential material accounting information that warrants auditors' further investigation. Consequently, application of the proposed method is expected to enhance the effectiveness of auditing practice. For investors, this method can be combined with the existing data mining techniques to construct better graphic mining techniques to form a more comprehensive examination of financial reporting quality for more effective firm valuation.

REFERENCES

- Bai, G., Hsu, S. H., & Krishnan, R. (2014). Accounting performance and capacity investment decisions: Evidence from California hospitals. *Decision Sciences, 45*(2), 309–339.
- Baldwin, A. A., & Brand, T. (2011). The impact of XBRL: A Delphi investigation. *International Journal of Digital Accounting Research, 11*, 1–24.
- Benbasat, I., & Dexter, A. S. (1986). An investigation of the effectiveness of color and graphical information presentation under varying time constraints. *MIS Quarterly, 10*(1), 59–83.
- Bhojraj, S., Lee, C., & Oler, D. K. (2003). What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research, 41*(5), 745–774.
- Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical Computer Science, 337*(1), 217–239.
- Boritz, J. E., & No, W. G. (2005). Security in XML-based financial reporting services on the internet. *Journal of Accounting and Public Policy, 24*(1), 11–35.
- Bradshaw, M. T., Miller, G. S., & Serafeim, G. (2009). Accounting method heterogeneity and analysts' forecasts. Unpublished paper, University of Chicago, University of Michigan, and Harvard University.
- Bunke, H., & Allermann, G. (1983). A metric on graphs for structural pattern recognition. *Proceedings of the 2nd European Signal Processing Conference EUSIPCO, 1983, Erlangen, Germany, 257–260*.
- Bushman, R. M., & Smith, A. J. (2001). Financial accounting information and corporate governance. *Journal of Accounting and Economics, 32*(1), 237–333.
- Carrillo, J. E., Druehl, C., & Hsuan, J. (2015). Introduction to innovation within and across borders: A review and future directions. *Decision Sciences, 46*(2), 225–265.
- Chan, L. K., Lakonishok, J., & Swaminathan, B. (2007). Industry classifications and return comovement. *Financial Analysts Journal, 63*(6), 56–70.

- Chen, W. (1998). More efficient algorithm for ordered tree inclusion. *Journal of Algorithms*, 26(2), 370–385.
- Chen, W. (2001). New algorithm for ordered tree-to-tree correction problem. *Journal of Algorithms*, 40(2), 135–158.
- Chong, D., & Zhu, H. H. (2012). Firm clustering based on financial statements. *Proceedings of 22nd Workshop on Information Technology and Information Systems (WITS'12), 2012, Orlando, Florida, US*.
- Chowdhuri, R., Yoon, V. Y., Redmond, R. T., & Etudo, U. O. (2014). Ontology based integration of XBRL filings for financial decision making. *Decision Support Systems*, 68, 64–76.
- Christensen, J. L. (2013). The ability of current statistical classifications to separate services and manufacturing. *Structural Change and Economic Dynamics*, 26, 47–60.
- Clarke, R. N. (1989). SICs as delineators of economic markets. *Journal of Business*, 62(1), 17–31.
- Clements, C. E., & Wolfe, C. J. (2000). Reporting financial results with the video medium: An experimental analysis. *Journal of Information Systems*, 14(2), 79–94.
- Cong, Y., Hao, J., & Zou, L. (2014). The impact of XBRL reporting on market efficiency. *Journal of Information Systems*, 28(2), 181–207.
- Dalziel, M. (2007). A systems-based approach to industry classification. *Research Policy*, 36(10), 1559–1574.
- De Franco, G., Kothari, S. P., & Verdi, R. S. (2011). The benefits of financial statement comparability. *Journal of Accounting Research*, 49(4), 895–931.
- Dhole, S., Lobo, G. J., Mishra, S., & Pal, A. M. (2015). Effects of the SEC's XBRL mandate on financial reporting comparability. *International Journal of Accounting Information Systems*, 19, 29–44.
- Fama, E. F., & French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43(2), 153–193.
- Fan, J. P., & Lang, L. H. (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. *Proceedings of the 34th International Conference on Information Systems, 2013, Milano, Italy*.
- Frederickson, J. R., Hodge, F. D., & Pratt, J. H. (2006). The evolution of stock option accounting: Disclosure, voluntary recognition, mandated recognition, and management disavowals. *The Accounting Review*, 81(5), 1073–1093.
- Frownfelter-Lohrke, C. (1998). The effects of differing information presentations of general purpose financial statements on users' decisions. *Journal of Information Systems*, 12(2), 99–107.
- Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: W. H. Freeman.

- Guenther, D. A., & Rosman, A. J. (1994). Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics*, 18(1), 115–128.
- Haque, W., Aravind, A., & Reddy, B. (2009). Pairwise sequence alignment algorithms: A survey. *Proceedings of the 2009 Conference on Information Science, Technology and Applications*, ACM, 96–103.
- Hirst, D. E., & Hopkins, P. E. (1998). Comprehensive income reporting and analysts' valuation judgments. *Journal of Accounting Research*, 36, 47–75.
- Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), 1423–1465.
- Hodge, F. D., Kennedy, J. J., & Maines, L. A. (2004). Does search-facilitating technology improve the transparency of financial reporting? *The Accounting Review*, 79(3), 687–703.
- Holthausen, R. W., & Leftwich, R. W. (1983). The economic consequences of accounting choice implications of costly contracting and monitoring. *Journal of Accounting and Economics*, 5, 77–117.
- 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.
- Jansson, J., & Lingas, A. (2001). A fast algorithm for optimal alignment between similar ordered trees. In *Combinatorial Pattern Matching*. New York: Springer, 232–240.
- Janvrin, D. J., Pinsker, R. E., & Mascha, M. F. (2013). XBRL-enabled, spreadsheet, or PDF? Factors influencing exclusive user choice of reporting technology. *Journal of Information Systems*, 27(2), 35–49.
- Jiang, T., Wang, L., & Zhang, K. (1995). Alignment of trees—An alternative to tree edit. *Theoretical Computer Science*, 143(1), 137–148.
- Kahle, K. M., & Walkling, R. A. (1996). The impact of industry classifications on financial research. *Journal of Financial and Quantitative Analysis*, 31(03), 309–335.
- King, D. R., & Slotegraaf, R. J. (2011). Industry implications of value creation and appropriation investment decisions. *Decision Sciences*, 42(2), 511–529.
- Klein, P. N. (1998). Computing the edit-distance between unrooted ordered trees. In *Algorithms—ESA '98*, Springer, 91–102.
- Knuth, D. E. (1998). *The art of computer programming: Sorting and searching*, Volume 3. Delhi, India: Pearson Education.
- Kolesnikoff, V. S. (1940). Standard classification of industries in the United States. *Journal of the American Statistical Association*, 35(209a), 65–73.
- Koonce, L., Lipe, M. G., & McNally, M. L. (2005). Judging the risk of financial instruments: Problems and potential remedies. *The Accounting Review*, 80(3), 871–895.

- Lee, C. M., Ma, P., & Wang, C. C. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, *116*(2), 410–431.
- Lenard, M. J., Alam, P., & Booth, D. (2000). An analysis of fuzzy clustering and a hybrid model for the auditor's going concern assessment. *Decision Sciences*, *31*(4), 861–884.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.
- Liu, C., Luo, X. R., Sia, C. L., O'farrell, G., & Teo, H. H. (2014). The impact of XBRL adoption in PR China. *Decision Support Systems*, *59*, 242–249.
- Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*(2), 129–137.
- Lymer, A. (1999). *Business reporting on the Internet*. London, UK: International Accounting Standards Committee London.
- Maines, L. A., & McDaniel, L. S. (2000). Effects of comprehensive-income characteristics on nonprofessional investors' judgments: The role of financial-statement presentation format. *Accounting Review*, *75*(2), 179–207.
- Miller, G. S., & Skinner, D. J. (2015). The evolving disclosure landscape: How changes in technology, the media, and capital markets are affecting disclosure. *Journal of Accounting Research*, *53*(2), 221–239.
- Narasimhan, R., Schoenherr, T., Jacobs, B. W., & Kim, M. K. (2015). The financial impact of fsc certification in the united states: A contingency perspective. *Decision Sciences*, *46*(3), 527–563.
- Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, *33*(1), 31–88.
- Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Proceedings of Advances in Neural Information Processing Systems Conference, 2001, Vancouver, Canada*, 849–856.
- Pagell, R. A., & Weaver, P. J. (1997). NAICS: NAFTA's industrial classification system. *Business Information Review*, *14*(1), 36–44.
- Perdana, A., Robb, A., & Rohde, F. (2014). An integrative review and synthesis of XBRL research in academic journals. *Journal of Information Systems*, *29*(1), 115–153.
- Pinsker, R. E., & Li, S. (2008). Costs and benefits of XBRL adoption: Early evidence. *Communications of the ACM*, *51*(3), 47–50.
- Rennekamp, K. (2012). Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research*, *50*(5), 1319–1354.
- Saeed, K. A., Malhotra, M. K., & Grover, V. (2011). Interorganizational system characteristics and supply chain integration: An empirical assessment. *Decision Sciences*, *42*(1), 7–42.
- Schipper, K. (2007). Required disclosures in financial reports. *The Accounting Review*, *82*(2), 301–326.

- Segars, A. H., & Grover, V. (1995). The industry-level impact of information technology: An empirical analysis of three industries. *Decision Sciences*, 26(3), 337–368.
- Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905.
- Stock, D., & Watson, C. J. (1984). Human judgment accuracy, multidimensional graphics, and humans versus models. *Journal of Accounting Research*, 22(1), 192–206.
- Tai, K.-C. (1979). The tree-to-tree correction problem. *Journal of the ACM*, 26(3), 422–433.
- Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *Journal of the ACM*, 23(1), 31–42.
- Vickery, S. K., Droge, C., & Markland, R. E. (1993). Production competence and business strategy: Do they affect business performance? *Decision Sciences*, 24(2), 435–456.
- Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.
- Yang, S., & Cogill, R. (2013). Balance sheet outlier detection using a graph similarity algorithm. *Proceedings of 2013 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, IEEE, 135–142.
- Yoon, H., Zo, H., & Ciganek, A. P. (2011). Does XBRL adoption reduce information asymmetry? *Journal of Business Research*, 64(2), 157–163.
- Zhang, K., & Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6), 1245–1262.
- Zopounidis, C., Doumpos, M., & Zanakis, S. (1999). Stock evaluation using a preference disaggregation methodology. *Decision Sciences*, 30(2), 313–336.

Steve Y. Yang is an assistant professor in the School of Business at Stevens Institute of Technology where he teaches graduate courses on stochastic calculus, algorithmic trading, and computational finance and economics. He received his PhD in Systems and Information Engineering from the University of Virginia. He currently serves as the Director of the Financial Engineering PhD program at Stevens. His research interests include market microstructure, behavioral finance, portfolio theory, and financial market simulation. His work has appeared in the *Journal of Banking and Finance*, *Quantitative Finance*, *Expert Systems with Applications*, and *Neurocomputing* among others.

Fang-Chun Liu is an assistant professor of accounting at Kate Tiedemann College of Business, University of South Florida-St. Petersburg. She received her PhD in Business Administration from Temple University. Her research interests include: strategic cost and performance management, corporate governance, human capital, executive compensation, financial reporting quality, and the financial implications

of information technology investment. Professor Liu has presented her research actively at the most prestigious international conferences in Accounting and Information Systems areas. Her work has published in *International Journal of Operations and Production Management*, *Information and Management*, and the *Proceedings of International Conference on Information Systems* among others.

Xiaodi Zhu is an assistant professor in the Department of Finance at New Jersey City University. She received her PhD degree in financial engineering from Stevens Institute of Technology. Her current research interests focus on finance and data analytics including topics such as behavioral finance, financial disclosure analysis, financial information efficiency, and portfolio analysis. Dr. Zhu has published and presented her research at various international conferences in financial information systems areas.

David C. Yen is currently a professor of MIS in the Department of Management, Marketing & Information Systems, SUNY-Oneonta. Professor Yen is active in research and has published books and articles which have appeared in *ACM Transaction of MIS*, *Decision Support Systems*, *Information & Management*, *International Journal of Electronic Commerce*, *ACM SIG Data Base*, *Information Sciences*, *Communications of the ACM*, *Government Information Quarterly*, *IEEE IT Professionals*, *Information Society*, *Omega*, *International Journal of Organizational Computing and Electronic Commerce*, and *Communications of AIS* among others. Professor Yen's research interests include data communications, electronic/mobile commerce, database, and systems analysis and design.