Information & Management xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

Industry classification with online resume big data: A design science approach

Xiaoying Xu[a], Hanlin Qian[a], Chunmian Ge[a,*], Zhijie Lin[b]

[a] School of Business Administration, South China University of Technology, 381 Wushan Road, Tianhe District, Guangzhou, 510000, PR China
[b] School of Economics and Management, Tsinghua University, Beijing, 100084, PR China

ARTICLE INFO

Keywords:
Industry classification
Labor mobility
Big data
Community detection

ABSTRACT

Industry classification is a vital step of industry analysis and competitive intelligence. However, existing schemes and methods are limited by the lagged information of firms' business and the lack of consideration of the human resource aspects. In this paper, we adopt a design science approach to develop and evaluate a novel industry classification method by constructing a labor mobility network using online resume big data collected from the professional social network. We also propose a hierarchical extension of the community detection algorithm to better discover scalable firm clusters on the constructed network. The evaluation conducted on real-world datasets shows that our method outperforms the existing industry classification schemes and the state-of-the-art methods by improving their explanatory power and enlarging the cross-industry variation. Moreover, two application cases confirm the validity of our method in earlier revealing firms' action of entering new industries.

1. Introduction

Competitive Intelligence (CI) is defined as "the process of developing actionable foresight regarding competitive dynamics and nonmarket factors that can be used to enhance competitive advantage" [1]. CI has grown during the past few decades to become a core capability of most large firms, and a survey shows that 55 percent of the Fortune 500 companies have decided to invest resources for developing and utilizing CI processes and products. A famous example of successful CI application is from IBM. After losing more than $14 billion during 1991–1993, IBM launched a pilot project of CI, which helped the company track and analyze its competitors intelligently, and this led to great success. After that, CI has become ingrained in IBM's corporate culture [2].

As pointed out by Prescott et al. [3], the primary objective of CI is providing an overview of industries and competitors, and the two secondary objectives include identifying the vulnerable areas of competitors and identifying competitors' potential moves that may endanger the firm's superiority. With regard to this, industry classification plays an important role in CI, as it is a vital step before investigating the focal firm's competitive environment and growth opportunities [4]. A good industry classification scheme should timely reflect the rapid changes of firms' business in the market space, which allows the management to respond promptly.

Two well-known industry classification schemes are the Standard Industrial Classification (SIC) and the North American Industry Classification System (NAICS), which are widely used in a wide range of business and management studies. Although well designed, there still exist a few limitations in the current industry classification schemes. The first limitation is the issue of timeliness. Given the rapid change of the modern economy, a firm may enter an industry at a faster pace than ever before. It is not strange that new industries may emerge as well. For example, according to the existing schemes, it is difficult to identify firms attributed to self-driving—an emerging new industry. As the current classification schemes are quite stable with relatively few changes across time, how to improve the timeliness of industry classification schemes becomes critical in an information-led and fast-changing economy. The second limitation is the lack of human-capital side information for industries. Human capital information is largely ignored in the current classification schemes, which mainly focus on the business or product side. This is understandable, as it is not easy to collect human capital information for firms in a systematic way. However, considering human capital as the most important asset and the carrier of both general and firm-specific knowledge [5], incorporating such information would be beneficial.

Recently, researchers have begun seeking solutions to supplement the industry classification schemes. For example, Hoberg and Phillips [6] and Fang et al. [4] performed industry classification by applying text mining to the 10-K forms. Their solutions based on annual reports

* Corresponding author.
E-mail addresses: bmxyxu@scut.edu.cn (X. Xu), 201620123470@mail.scut.edu.cn (H. Qian), bmgecm@scut.edu.cn (C. Ge), linzhj@sem.tsinghua.edu.cn (Z. Lin).

https://doi.org/10.1016/j.im.2019.103182
Received 1 September 2018; Received in revised form 23 July 2019; Accepted 26 July 2019

such as 10-K forms have shown some advantages but still cannot timely reflect the dynamics of the firms' business due to the low update frequency and the lagged information. The reason is that firms like Google would only update their new business on self-driving in the annual reports long after they have entered this field but not at the early stage. In addition, these industry classification solutions still fail to consider the human capital aspects.

Fortunately, with the rapid development of information technology, individual data are getting increasingly available on the Internet and have been intensively used by researchers [7]. With this in mind, we aim to employ a Design Science approach [8,9] to address these two limitations by utilizing the online labor mobility big data. The IT artifact created by this work would be an automatic industry classification architecture, which would be able to leverage the human capital information embedded in the online professional social network and which could reflect the industry changes in a timelier way. It should be noted that the purpose of this paper is not to come up with a method that can completely replace the existing ones. Instead, the proposed solution is intended to be used as a supplement of extant industry classification approaches by starting an investigation from a new perspective, that is, the labor mobility, which has been rarely concerned by existing research.

Specifically, we propose a novel solution based on the fact: when a firm decides to enter a new field, it must first recruit people in the relevant area or with relevant skills. Guided by the works [10] and [5], which contribute to Resource-Based View by analyzing industry groups from a human-resource-related perspective, we propose a brand-new industry classification method by constructing and analyzing the labor mobility network and use this network to perform industry analysis. To achieve this goal, in detail, we collect a rich dataset of online resumes from a major professional social network website and extract the turnover records of each individual from their online resumes, with which we construct the labor mobility network among different firms.

To discover the clusters of firms that consist of the latent industry classes, we adapt and extend a community detection algorithm, that is, the Louvain algorithm [11], to a hierarchical version and then apply it on the constructed network. Next, we measure turnover records between firms in different industries to find unusual connections to detect potential actions of industry entries. We follow the Design Science Research principles and evaluate the proposed methods rigorously. The evaluation results based on real-world datasets show that the proposed method outperforms other commonly used schemes and form-10 K-based methods in terms of explaining stock return variation and enlarging across-industry variation. Moreover, the proposed method has been shown to be able to reveal a firm's action of entering a new industry.

This paper designs and implements a brand-new solution for automatic industry classification. By leveraging labor mobility big data from the online professional social network, we propose an industry classification architecture that can dynamically reflect the entry/leave of firms in a timely manner. This addresses a major limitation of the current industry classification schemes. Additionally, the proposed architecture incorporates the human capital information of firms, which is rarely considered by extant research, helping address the other limitations of the current schemes. Further, our approach of constructing and analyzing the labor mobility network has strong implications for both researchers and practitioners in related fields.

The remaining parts of this paper are organized as follows: first, we review related research. Then we describe the intuition behind our proposed method and provide an overview, followed by a more detailed elaboration. Further, we evaluate the method and present the evaluation results. Lastly, we summarize our work and discuss the possibilities for future work.

2. Related work

In the following sections, we will review related work on industry classification, including its latest extensions. Then we will demonstrate related research of Resource-Based View. In addition, we will discuss related research and applications on online professional social network. Methods for community detection will also be introduced, which will be adapted in our approach for industry classification.

2.1. Industry classification

There are three traditional industry classification schemes: Standard Industrial Classification (SIC), North American Industry Classification System (NAICS), and Global Industry Classification Standard (GICS). The American government first developed the SIC in 1937 and later replaced it with the NAICS in concert with Canadian and Mexican governments. Additionally, Standard & Poor's and Morgen Stanley Capital International established the GICS that is based on the judgment of financial analysts to determine which firms are financially comparable. All these government industry classification schemes employ actively based approaches, which define an industry as "a generally homogeneous group of economic producing units, primarily engaged in a specific set of activities" and define an activity as "a particular method of combining goods and service inputs, labor and capital to produce one or more goods and/or services (products)." Such schemes refer to the nature of product markets in terms of seller and buyer concentration level, product differentiation, and barriers to entry, instead of the relationships between firms [12].

Although widely used, the existing schemes are limited by some well-known drawbacks. First, because of the small number of industry categories, they may fail to reflect the real industrial structure and therefore lead to misclassification. For example, Dalziel [12] states the example that the firms in the communications equipment subsector cannot be identified by considering 20 NAICS sectors. Furthermore, the low update frequency would prevent them from capturing the evolvement of firms' business and the change of industry structure [13]. For example, the first version of NAICS was released in 1997, and after that, only three revisions were released, in 2002, 2012, and 2017. Lastly, Fang et al. [4] point out that these schemes assume a binary relationship and cannot measure the degree of relatedness between two firms.

Noticing these limitations, some researchers have shifted their focus to classify industries on different aspects. For example, Fang et al. [4] propose to extract topic features from business descriptions of form 10-K for industry classification. Similarly, Hoberg and Phillips [6] also use the business descriptions from form 10-K to calculate firms' product similarities for industry classification. Chong and Zhu [13] introduce a firm clustering method that employs XBRL-based financial information. These new approaches work well by revealing more dimensions of the industry structure by utilizing the textural content of firms' annual reports. However, they still lack timeliness caused by the lagged information from the data they use.

2.2. Resource-based view

There has been some theoretical research of industry analysis that employs the Resource-Based View [14] and lays the theoretical foundation of our work. The Resource-Based View is a managerial framework that has been widely applied to determine the strategic resources having the potential to bring comparative advantage to a firm. It proposes that each firm is heterogeneous because of its heterogeneous resources.

The work of Farjoun [10] proposes the Resource-Related Industry Groups (Resource-RIGs) identified by the similarities in the

requirements for human expertise, which has provided a theoretical grounding for the reasonability of industry classification from a human resource perspective. In addition, the work of Neffke and Henning [5] also highlights that the similarities in human capital, or skill relatedness, are extremely useful for characterizing industries. However, the human capital information of firms has been largely overlooked by extant industry classification research. To bridge this research gap, our proposed method considers the human resource aspect and makes use of the labor mobility big data from online professional social network that will be introduced below.

2.3. Online professional social network

Online professional social network such as LinkedIn, Viadeo, and Open Science Lab is a type of social network focusing exclusively on business relationships and interactions [15]. Many studies have been conducted with data from such networks. For example, Ge, Huang, and Png [16] use LinkedIn data to investigate the effect of human capital on mobility of engineers and scientists and find that LinkedIn provides more accurate career histories than patents through an inventor survey. Antoine, Cécile, and Hilda [17] demonstrate that LinkedIn allows decomposing firms' value chains and permits to develop interlocking networks dedicated to firms' divisions. Wang, Li, and Zhou [18] implement personal profile summarization by leveraging both textual information and social connection information from LinkedIn. These existing studies have demonstrated that online professional social network is a valuable and reliable source for investigating the human-resource-related problems. However, no existing work has been found to utilize the online professional social network data for industry analysis, which is the focus of this paper.

2.4. Community detection

Community detection in network is one of the hottest topics in modern network science [19]. Communities, also called clusters or modules, are groups of vertices that are more likely to be connected to each other than to vertices in other groups [20]. The methods for finding communities in a specific network are called community detection. Compared with traditional clustering methods, community detection is more suitable for large-scale and rapid-change situation [21], making it appropriate for us to discover firm clusters. Community detection algorithms have been widely used in many areas. However, scant attention has been drawn to applying community detection to industry analysis. In an extensive search of the literature, we were unable to find work that addresses competitive intelligence by adapting these community modeling techniques. Therefore, our method aims to generate scalable firm groups by adapting and extending the community detection algorithm.

Existing community detection research can be classified into two streams. One stream of research aims at maximizing the modularity of the detected communities, such as Louvain [11], Agglomerative Greedy [22], Modularity Optimality [23], and Combo Optimization [24]. Another stream of research employs the Random Work model. The underlying intuition is that in a random walk, the probability of remaining inside a community is higher than that going outside due to the higher density of the internal edges. Such algorithms include Walktrap [25], Infomap [26], and Relaxmap [27].

As the labor mobility network may contain a huge number of nodes and edges, the selected algorithm in this study should be highly scalable. Sobolevsky et al. [24] and Bae et al. [28] compare the performance of these existing algorithms, and their results show that the Louvain algorithm is among the best in terms of quality and scalability. In addition, the Louvain algorithm has relatively simple procedures, making it feasible for us to derive a hierarchical extension that enables multi-level analysis.

3. Intuition and overview

We are interested in helping market practitioners and researchers with a better industry classification approach that is able to reveal more aspects of firms' business and response quickly to market changes. Our proposed method is guided by the theory of Resource-Based View and is motivated by a simple fact that when a firm starts a new business, it must recruit people in relevant areas. Hence, from the human resource the firm owns or needs, we may infer its business. Furthermore, people with expertise in a certain area would tend to stay among the firms related to the same area. For instance, a person who works in Google is more likely to take an offer from Apple than from McDonald's. With this nature, it is reasonable to expect that in a network representing the labor mobility among different firms, those firms focusing on similar business would be well connected with each other due to the high mobility of laborers between them. We call such network as a labor mobility network, in which a node indicates a firm, and a weighted directed edge between two nodes indicates the number of people who leave one firm and join another. Based on the network structure, our method is designed to classify the firms on the labor mobility network into subgroups that consist of latent industry classes.

As the industry classes are self-generated rather than a small number of predefined categories, our method would be able to cover more aspects of firms' business and industry structure, contributing to more precise industry classification. Moreover, we make use of the human resource information of a firm to capture its changes in business, which can quickly respond to market changes, addressing the lagging problem and therefore improving the precision of industry classification.

4. Solution details

In this section, we will first describe the architecture of our proposed solution, followed by the details of each component in the architecture. The system architecture of our approach is depicted in Fig. 1. There are three major components in our architecture: Resume

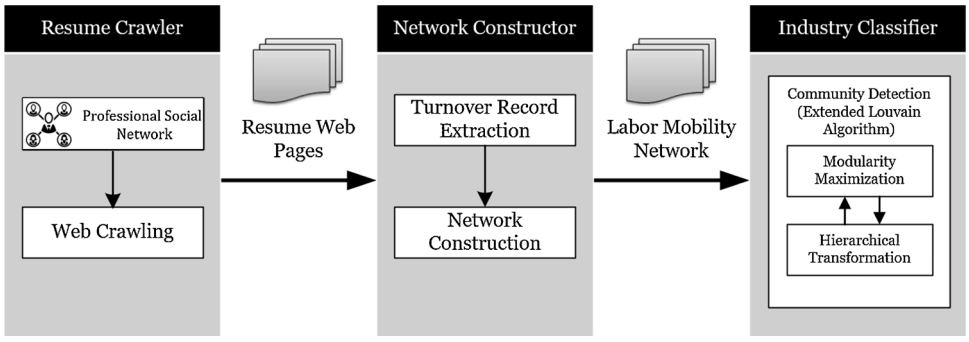

Fig. 1. Industry classification architecture.

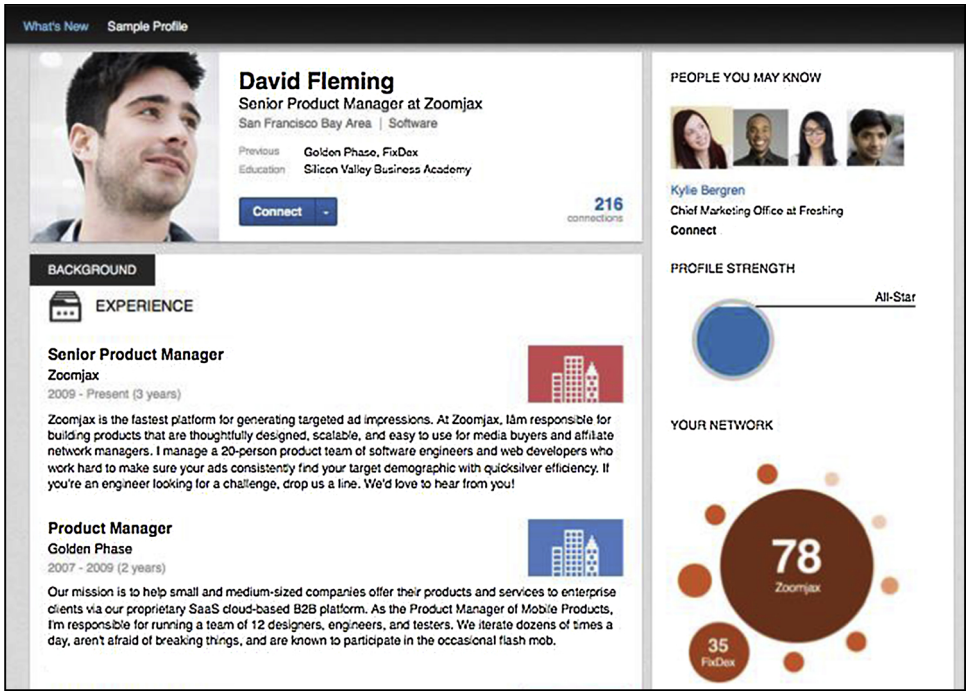

Fig. 2. Resume webpage on the professional social network.

Crawler, Network Constructor, and Industry Classifier. More details of each component will be discussed in the ensuing sub-sections.

4.1. Resume crawler

For the purpose of research, we develop a Resume Crawler to collect a sample of public profiles from the professional social network for validating the effectiveness of the proposed method. However, for the potential users of the designed artifact, for example, large enterprises, and business analytics and consulting firms, it is also feasible to obtain the required data through cooperation with or purchasing paid data service from professional social network platforms.

The main task of the Resume Crawler is to collect the online resumes of individuals from the professional social network for further analysis. The crawler integrated with the Google API is able to systematically and automatically browse the professional social network website and download public resume webpages. Fig. 2 shows a sample of online resume with working experience and other personal information. During the process of data collection, all non-job-related content like advertisements is filtered and excluded. The collected data are then passed to the next processing component.

4.2. Network constructor

This component is designed to construct the labor mobility network, an important element in our implementation. There are two main steps to implement this component:

4.2.1. Turnover record extraction

The turnover records are embedded in the working experience section in each webpage of individual resume. They need to be extracted and reorganized. Specifically, we apply a commonly used html parser JSoup, which is a Java library for extracting content from webpages, to extract the working experience content from each resume webpage. The extracted working records are stored in the database, and then they are reorganized into every turnover record represented as a tuple (*employee ID, firm_from, firm_to, year*) by analyzing the start-time and end-time of each working record. For example, in the resume webpage shown in Fig. 2, we can see that David left Golden Phase and joined Zoomjax in 2009, so a turnover record "Golden Phase - > Zoomjax, 2009" can be extracted from this resume, which can be represented as a tuple (employee1, Golden Phase, Zoomjax, 2009).

4.2.2. Network Construction

With the extracted turnover records at the individual level, this step is to aggregate these records at the firm level and to construct the labor mobility network that represents the mobility of laborers among different firms. With each individual turnover tuple (*employee ID, firm_from, firm_to, year*), a firm-level aggregation can be also represented as a tuple (*firm_from, firm_to, number_of_employees, year*). Then, a labor mobility network can be constructed. Let $G_{year} = (V, E)$ be a directed graph with a set of vertices V and a set of directed edges E in a given year. A vertex V_i denotes a firm i (i.e., *firm_from*), and a directed edge $E_{i,j}(\rho)$ denotes that there were ρ employees (i.e., *number_of_employees*) who left firm i and joined firm j (i.e., *firm_to*). Fig. 3 shows a simple example of a labor mobility network. In this example, a tuple (Microsoft, Apple, 50, 2009) can be converted into a directed edge from vertex *Microsoft* to vertex *Apple* weighted by 50, indicating that there were 50 employees who left Microsoft and joined Apple during the year 2009.

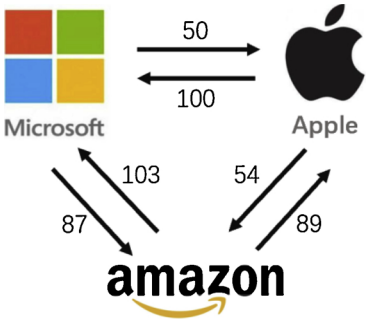

Fig. 3. Example of a labor mobility network.

4.3. Industry classifier

The industry classifier is a core component to discover the latent industry classes and perform industry classification. Generally, with the labor mobility network constructed by the previous component as the input, this component is used to analyze the closeness between vertices representing firms by looking at the network structure and to form communities of vertices, which can denote industries. To fulfill this task, we extend an algorithm called Louvain [11], which is acknowledged as one of the most excellent and fastest community detection algorithms, under our context. This process consists of two main steps:

4.3.1. Modularity maximization

The objective of the Louvain algorithm is maximizing the modularity Q that is defined as

$$Q = \frac{1}{2\sum Emp}\sum_{c\in C}\left[\sum_{f \text{ and } t\in c} Emp - \frac{(\sum_{f \text{ or } t\notin c} Emp)^2}{2\sum Emp}\right] \quad (1)$$

With the tuple (*firm_from, firm_to, number_of_employees, year*) that represents labor mobility from one firm to another, *Emp* denotes the *number_of_employee*; $\sum Emp$ denotes the total number of laborers moving on this network; f denotes the *firm_from*; t denotes the *firm_to*; C denotes the collection of all communities; and c denotes a specific community in C. Modularity measures whether a particular division of the target network is meaningful and natural [29]. Intuitively, a high modularity indicates that the level of closeness (measured by labor mobility) between firms from the same community is high, and the level of closeness between firms from different communities is low.

Fig. 4 depicts the procedures of the Louvain algorithm. Generally, the original algorithm is implemented in the following steps. Initially, we consider every firm as an independent industry. Then we try assigning a firm to every other industry and calculate the changes of modularity in each trial. This firm is allocated to the target industry where the largest modularity gain is obtained. We iterate this industry assignment process for other firms until the algorithm converges, which means there is no gain in modularity.

4.3.2. Hierarchical transformation

The original community detection algorithm tends to discover small communities, resulting in too many firm clusters, which decreases the usefulness of the results. For example, in a network with 100,000 firms, the Louvain algorithm may detect approximately 10,000 communities. To get desirable results with appropriate scale, we propose a hierarchical extension to the original Louvain algorithm. The extended algorithm combines the nodes in the same community given by the previous step into new nodes and uses the combined nodes as the input to run the Modularity Maximization process. We iterate this loop until the number of communities reaches a desired scale, which can be specified in real-time applications. In our implementation, we set this number to 20 to compare with other schemes and methods.

Fig. 5 shows a simple example of two loops, which generate two levels of outputs. Specifically, the first output of the Louvain algorithm gives the lowest level of detected communities (Step 1 in Fig. 5). We then combine all firms in the same community into a new vertex and form a higher level of labor mobility network (Step 2 in Fig. 5), which can be represented as tuples (*community_from, community_to, number_of_emplyees, year*), where *number_of_emplyees* denotes the number of laborers move from one community to another. The Louvain algorithm can be executed again using these new tuples as the input to identify high-level communities (Step 3 in Fig. 5). In this step, the modularity Q' to be maximized can be defined as

$$Q' = \frac{1}{2\sum Emp'}\sum_{c'\in C'}\left[\sum_{f' \text{ and } t'\in c'} Emp' - \frac{(\sum_{f' \text{ or } t'\notin c'} Emp')^2}{2\sum Emp'}\right] \quad (2)$$

where *Emp'* denotes the *number_of_employee* in the tuple (*community_from, community_to, number_of_emplyees, year*); $\sum Emp'$ denotes the total number of laborers moving on the higher level network; f' denotes the *community_from*; t' denotes the *community_to*; C' denotes the collection of all second-level communities; and c' denotes a specific community in C'.

4.4. Evaluation

To evaluate the effectiveness of our proposed method, we implemented a prototype of our design artifact. Specifically, the Resume Crawler and Network Constructor were implemented in Java, and the Industry Classifier was implemented in R with the *igraph* and *dplyr* packages. The data were stored in MySQL database. The experiments were run in a desktop with Windows 10 OS, Intel Core i5-7200U CPU, and 8GB memory.

In this section, we report the results of the experiments we conduct to evaluate the effectiveness of our proposed method. First, we describe the dataset we use for the evaluation. Then we introduce our evaluation metrics, followed by the evaluation results of network analysis, qualitative results of identified industries, the comparison results with state-of-the-art, as well as two cases of application.

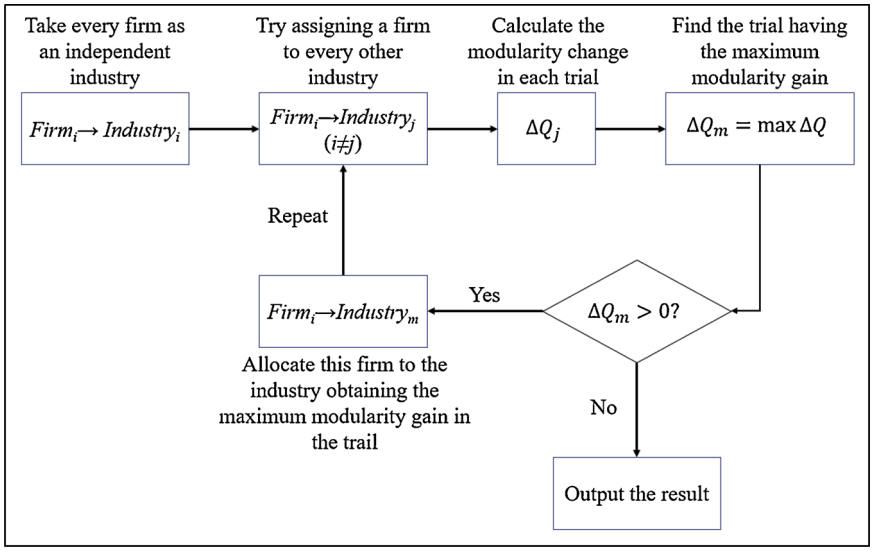

Fig. 4. Procedures of the louvain algorithm.

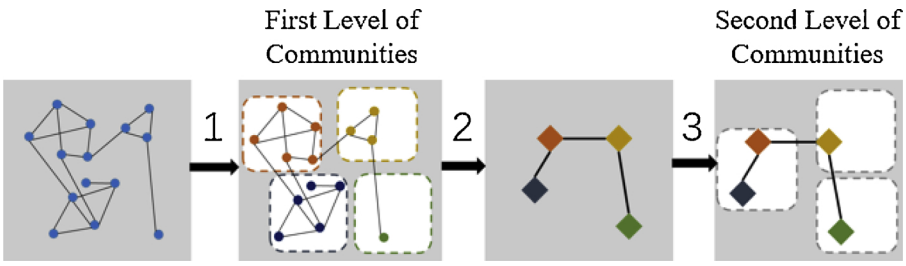

Fig. 5. Hierarchical community detection.

Table 1
Dataset statistics.

Year	2011	2012	2013
No. of Firms	194530	189435	199893
No. of Public Firms	4455	4265	4250
No. of Matched Firms	4239	4123	4067

4.5. Data collection

We collect 1.52 million online resumes from a major online professional social network, containing 5.74 million working records and 2.83 million turnover records for constructing the labor mobility network. The professional social network platform recognizes the names of the public firms and provides the corresponding tickers for the public firms, which allows us to match the financial and marketing data of the firms in the Compustat database. The numbers of public firms and matched records in the Compustat database on a yearly basis are reported in Table 1.

To compare our method with other existing methods, following Hoberg and Phillips [6] and Fang et al. [4], we choose the constituents of the Standard and Poor's 1500 (S&P 1500) as our firm sample, which is a combination of the S&P 500, the S&P MidCap 400, and the S&P SmallCap 600 and covers approximately 90% of market capitalization in the U.S., and collect the corresponding capitalization information from Yahoo Finance.

4.6. Evaluation metrics

4.6.1. Metrics for network analysis

The purpose of network analysis is to take a general view of the network structure and to justify whether the network is suitable for further analysis. We choose three typical indicators to analyze our network structure: *Density, Modularity*, and *Clustering Coefficient*.

Density of a network is defined as the number of connections a network has, divided by the total possible connections the network could have

$$D = \frac{2\,|E|}{|V|(|V|-1)} \tag{3}$$

where $|E|$ and $|V|$ denote the total number of edges and vertices, respectively, in a network. A higher density indicates that there is higher labor mobility among different firms, providing more useful information for detecting industries.

Modularity defined in Eq. 1 measures the strength of division of a network into sub-modules. Similar to Modularity, *Clustering Coefficient* is another measure of the degree to which nodes in a network tend to cluster together:

$$C = \frac{3\tau_\Delta}{\tau_3} \tag{4}$$

where τ_Δ is the number of triangles and τ_3 is the number of connected triplets of vertices in a network. We calculate the average Modularity and average Clustering Coefficient for all detected industries. A higher average Modularity and a higher average Clustering Coefficient indicate that the firms are closer to each other in the same industry than to the firms in other industries; thus, the boundaries of the identified industries are clearer. They also imply that the labor mobility network is more suitable for industry analysis from the perspective of the network structure.

4.6.2. Metrics for industry classification

Following Bhojraj et al. [30], Hoberg and Phillips [6], and Fang et al. [4], we compare our proposed industry classification method with other methods and commonly used industry classification schemes by measuring how they can help to explain the movements in target firms' stock returns. The validity of using such measurement has been also confirmed by the study of Lamponi [31], which finds the data-driven categorizations of large U.S. stocks show a consistent overlap with industries. Thus, a good industry classification scheme should have stronger explanatory power for the stock price co-movements. We will estimate the regression specification defined in the following equation:

$$R_{i,t} = \alpha_t + \beta_t R_{p,t} + \varepsilon_{i,t} \tag{5}$$

where $R_{i,t}$ denotes the quarterly return of firm i and $R_{p,t}$ denotes the average quarterly portfolio return of the industry p which firm i belongs to. Following Hoberg and Phillips [6], we pick 10 firms in a portfolio and run cross-sectional regressions between 2011 and 2013 on a monthly basis. Based on 36 regressions, we can obtain an average adjusted R^2, which can be used as a measurement of the effectiveness of industry classification: the higher the value of the average adjusted R^2, the more effective is the industry classification scheme.

Additionally, Hoberg and Phillips [6] propose using the cross-industry variation to evaluate the industry classification outcomes, based on the conclusion that a classification result is more informative if it generates a higher degree of cross-industry variation. We also adopt such metric in our evaluation. Specifically, for each detected firm group, we calculate the average characteristics of the firms inside it:

$$\sigma_f = \sqrt{\frac{\sum (\bar{f}_i - u)}{NI}} \tag{6}$$

where f is a firm characteristic, such as asset, market beta, etc.; $\bar{f}_i$ is the average characteristic of all firms in an industry; u is the overall average characteristic of all industries; and NI is the total number of industries.

4.7. Evaluation results

In this section, we first present the results of network analysis and identified industries and then report the evaluation results on industry classification compared with other methods. Lastly, we show two cases to demonstrate how the proposed method can reveal potential company actions of entering a new field.

4.7.1. Results of network analysis

Although we can construct the labor mobility network in any time span, to make our method comparable with other existing methods, we construct the network on a yearly basis. The descriptive results on the network analysis are shown in Table 2.

Table 2
Results on network analysis.

Year	2011	2012	2013
Vertex Scale	194530	189435	199893
Modularity	0.582	0.684	0.600
Density	0.312	0.359	0.318
Clustering Coefficient	0.647	0.702	0.680
Central companies	Google Amazon Apple Disney	Apple Google Disney Microsoft	Apple Google Pixar Intel Adobe Intuit

Table 3
Top 5 identified industries[1] .

Industry	Label	Representative firms	No. of firms
1	Internet	Apple Microsoft Alphabet	23456
2	Finance	Bank of America Citigroup Morgan Stanley	19658
3	Manufacturing	Hewlett-Packard Whirlpool Sony	19214
4	Fast Moving Consumer Goods	Coca-Cola Pepsi Americas Boston Beer	14484
5	Retailing	Wal-Mart Stores Costco Wholesale Dicks Sporting Goods	13112

From the indicator of *Density*, we notice that the network is relatively dense, given the fact that the highest density in the real social networks has been found to be 0.5 [32]. It implies that the constructed labor mobility network contains abundant information for further analysis. The *Modularity* values are all above the threshold 0.3 recommend by Newman [29], and the *Clustering Coefficient* values are all above 0.5, which is considered high in practice. They indicate that there are potential clusters in the constructed network, and the nodes inside each potential cluster would be tightly connected with each other. We can also see that there are some changes in the central companies during these 3 years, which means the network is dynamic but not stable. This is consistent with our expectation that the labor mobility network could be aligned with the market changes. Overall, the results of network analysis have shown that it is feasible to proceed to perform community detection.

4.7.2. Results of identified industries

A qualitative assessment is conducted based on the data in the year 2012 to show more intuitive results of industry classification. To make the results more interpretable, we asked 4Ph.D. students to label the 20 identified firm clusters. These students are majoring in management and have general knowledge of firms and industries. Each student was asked to label every identified firm cluster by choosing a unique industry label from a predefined label set that was constructed by selecting 30 most relevant labels from the NAICS codes. From the collected results, we find that the label of each firm cluster can be uniquely determined by using the one chosen by most annotators. To measure the inter-annotator agreement when labeling the firm clusters, we calculate the Fleiss' kappa, which is a statistical measure for assessing the reliability of agreement between more than 2 annotators. In our case, the value of Fleiss' kappa is 0.7396, indicating “substantial agreement” according to Landis and Koch [33].

[1] The full list of firms in all identified industries can be found in https://github.com/INFMAN2018696/Industry-Classification-List/blob/master/public_companies.xlsx.

Table 3 reports the top 5 industries having the largest number of firms identified by the proposed method and lists 3 representative firms in each industry. From the results, we can see that the identified industries are interpretable and reasonable, which confirm the validity of our proposed method. Similar results can be obtained with the data in the year 2011 and 2013. For the sake of space, they are not shown again.

4.7.3. Results of explaining stock returns

To show the effectiveness of our method based on the labor mobility network (denoted as LMN), we compare the average adjusted R^2 of the proposed LMN method with the commonly used schemes, that is, SIC, NAICS, and GICS. We also compare the proposed LMN method with the HP method [6] and the FDD method [4], both of which are based on the textual descriptions of 10 K forms, representing the state-of-the-art. In addition, other two baseline community detection algorithms, that is, INM (Infomap [26]) and MO (Modularity Optimality [23]) is also included to examine the validity of algorithm selection. As we have difficulties in replicating the HP and FDD methods, we directly use the results reported in the original papers. To reduce the differences in the number of industry groups, following Bhojraj et al. [30], we use the two-digit SIC codes, first three-digit NAICS codes and six-digit GICS codes. For the proposed LMN method, we iterate the community detection algorithm until the number of communities is comparable with the number of industry groups of other methods and is easy to interpret, which is 20. Although the differences in the number of industry groups remain, Bhojraj et al. [30] have shown that such differences would not have significant impact on the results. The results are shown in Fig. 6, and the percentage of improvement of our method on other methods is shown in Table 4.

The result shows that the average adjusted R^2 values of our LMN method are higher than those of all the other methods. Specifically, compared with the traditional schemes, our LMN method improves the average adjusted R^2 of the best one, that is, GICS, by 8.31% and 4.67% on the S&P 500 and S&P 1500 samples, respectively. When it comes to the state-of-the-art, our method outperforms the FDD method, which is the better one by improving its average adjusted R^2 by 1.23% and 3.97% on these two samples, respectively. We also notice that the magnitude of improvement is relatively small. A possible explanation is that the stock returns may be affected by many factors, and the

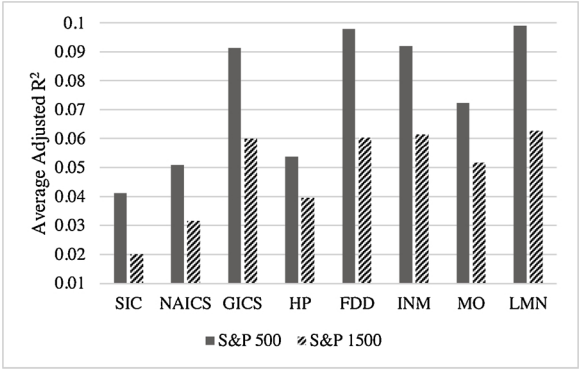

Fig. 6. Comparison on average adjusted R^2.

Table 4
Percentage of improvement of average adjusted R^2.

	SIC	NAICS	GICS	HP	FDD	INM	MO
S&P 500	140.53%	94.31%	8.31%	84.20%	1.23%	7.60%	36.88%
S&P 1500	212.43%	98.73%	4.67%	58.59%	3.97%	2.11%	21.47%

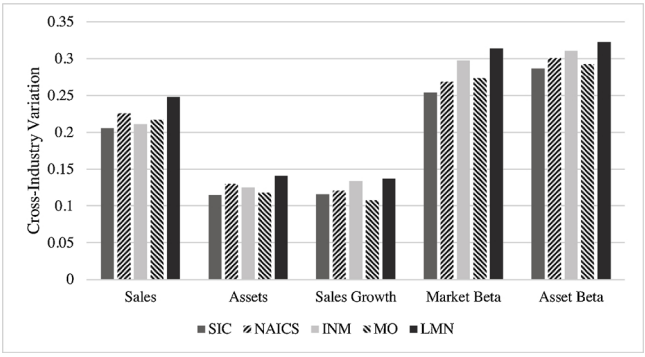

Fig. 7. Comparison on cross-industry variation.

industrial effect is only a very small part. When the explanatory power of the industrial effect is close to the upper bound, it is hard to be significantly improved.

With regard to algorithm selection, we can see that the proposed LMN method based on the Louvain algorithm outperforms the other two community detection algorithms, INM and MO, and increases their explanatory power by 7.60% and 36.88% on the S&P 500 sample and by 2.11% and 21.47% on the S&P 1500 sample, respectively. The results confirm the validity of our community detection algorithm selection.

4.7.4. Results of cross-industry variation

To further examine the validity of our method, we also adopt the evaluation approach proposed by Hoberg and Phillips [6] and compare the cross-industry variation given by our method in terms of sales, assets, sales growth, market beta, and asset beta, respectively, with the SIC, NAICS schemes, and the other two community detection algorithms, INM and MO. As the original papers of the HP and FDD methods do not inspect the cross-industry variation, and we have difficulties in replicating these two methods, they are not included in this comparison. The results are depicted in Fig. 7, and the percentage of improvement of our method on the other methods is given in Table 5.

Comparing the results, it can be seen that the industries classified by our proposed LMN method have higher cross-industry variation than both the SIC and NAICS schemes in terms of all firm characteristics, and the improvements are relatively significant. In detail, the LMN method increases by at least 7.31% of the cross-industry variation of the commonly used schemes. The results confirm that our proposed LMN method is able to provide a more informative industry classification. Further, the results confirm the superiority of using the extended Louvain algorithm by showing higher cross-industry variation in all financial aspects than the other two community detection algorithms.

Table 5
Percentage of improvement of cross-industry variation.

	SIC	NAICS	INM	MO
Sales	20.39%	9.73%	17.54%	14.29%
Assets	22.61%	8.46%	12.80%	19.49%
Sales Growth	18.10%	13.22%	2.24%	26.85%
Market Beta	23.62%	16.73%	5.37%	14.60%
Asset Beta	12.54%	7.31%	3.86%	10.24%

4.7.5. Cases of revealing firm's new industry entry

In addition to improving the explanatory power of the industrial effect, our method is designed in a hierarchical manner that allows revealing the industry structure at different scales, which is not supported by most community detection methods. In addition, the proposed method investigates the industry structure from the perspective of human resource that may provide useful information ahead of the annual reports and other industry classification schemes.

We use two cases to demonstrate the possible practical application of our method in revealing a firm's action of entering a new industry. A partial labor mobility network in 2011 constructed from the online resumes is depicted in Fig. 8. Nodes with different colors represent different industries detected by our method. The community on the top left is more related to the industry "Internet," the community in the middle denotes the industry more related to "Electronics" and the remaining community is more related to "Automobile."

As the first case, it is reasonable to see that Google (denoted as a red node in Fig. 8) is allocated among the Internet firms. It is interesting to find that there are some connections between Google and the "Automobile" community, from which we may infer that Google was entering the area of self-driving. In fact, Google's self-driving project began around the year 2010, and beta tests were conducted in 2012. Only after that, some information about the self-driving project was disclosed. However, neither Google nor its holding company Alphabet mentioned this self-driving project in form 10-K until 2017. Hence, this case shows that our method is able to reveal a firm's action of entering a new field ahead of other information channels, which provides great practical potentials.

Microsoft can be the second case. Similarly, Microsoft (denoted as a green node in Fig. 8) is attributed to the "Internet" industry. It is not surprising to see some connections between Microsoft and the "Electronics" community because Microsoft launched the project of Surface, which is the first personal computer designed and developed by itself, 3 years before 2012 when the first Surface was released. The project brought the urgent need for laborers in the field of electronics, therefore allowing us to observe such interesting connections. However, only after 2012 can we find the announcement of the Surface project in the firm's form 10-K. Again, this case confirms the superiority of our proposed method in discovering a firm's new industry entry.

5. Conclusion

In this paper, we adopt a design science approach to come up with an innovative industry classification method with the aim of addressing the problems caused by the lagged information of firms' business and the ignorance of human resource information. To achieve this goal, we employ the Resource-Based View as our theoretical foundation and propose to classify firms from the perspective of human resource. We design and implement an automatic industry classification solution. Our method extracts the turnover records from online resumes on the professional social network for constructing the labor mobility network. We also extend the community detection algorithm to a hierarchical version, making it more suitable for our purpose of discovering latent firm clusters. The evaluation has confirmed the validity of our design.

Our work contributes to Resource-Based View by empirically testing the effectiveness of using human resource information in industry analysis. Based on the analysis in this work, we can conclude that human resource information is an essential element in industry classification. By taking the perspective of labor mobility, the proposed method is possible to capture minor industry changes with a finer granularity and in a timelier manner, which differentiates itself from other approaches. Our novel approach of constructing the labor mobility network and the proposed extension of hierarchical community detection have both theoretical and practical implications for related fields.

There are a few limitations in our research. First, our proposed

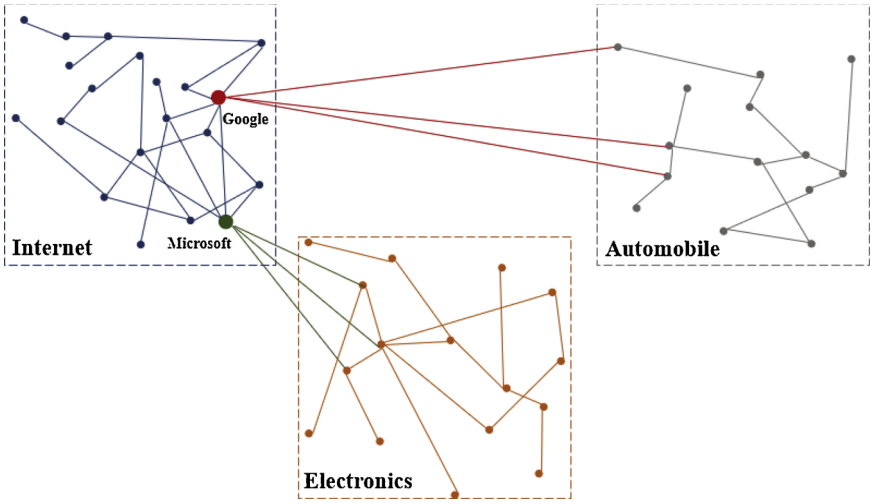

Fig. 8. A Partial Network in 2011.

method is not able to detect a firm's business changes or its new industry entry if the firm can realize such actions with existing employees and does not need to recruit external laborers. Such limitation can be alleviated by integrating with other industry classification schemes. Second, the professional social network users in our sample are not evenly distributed among all industries. Generally, the IT industry tends to have more users. This sample bias problem could be addressed in future work by integrating other data sources. Third, the timeliness of our method in reflecting market changes has not been quantitatively tested. It is possible to come up with suitable metrics for quantifying and evaluating the timeliness in future work. Lastly, in future work, it would be interesting to apply our method in investigating the impact of intra-industry labor mobility and inter-industry labor mobility on the performance of firms, which has not been discussed in this work.

Acknowledgments

We gratefully acknowledge the funding support from the National Natural Science Foundation of China (Grant Nos. 71601081, 71872065, 71503084, 71832003, 71872080 and 71502079) and Guangdong Natural Science Foundation (Grant Nos. 2016A030310426 and 2017A030312001), Young Scholars of Guangzhou (Grant No. 18QNXR05) and the Fundamental Research Funds for the Central Universities of the South China University of Technology (Grant No. XYMS201902).

References

[1] J. Prescott, The evolution of competitive intelligence, Int. Rev. Strateg. Manage. 6 (1999) 71–90.
[2] L. Behnke, P. Slayton, Shaping a corporate competitive intelligence function at IBM, Competitive Intell. Rev. 9 (1998) 4–9, https://doi.org/10.1002/(SICI)1520-6386(199804/06)9:2<4::AID-CIR3>3.0.CO;2-E.
[3] J.E. Prescott, G. Bhardwaj, Competitive intelligence practices: a survey, Competitive Intell. Rev. 6 (1995) 4–14, https://doi.org/10.1002/cir.3880060204.
[4] F. Fang, K. Dutta, A. Datta, LDA-based industry classification, Thirty Fourth International Conference on Information Systems 3 (2013) 2500–2509.
[5] F. Neffke, M. Henning, Skill relatedness and firm diversification, Strateg. Manage. J. 34 (2013) 297–316, https://doi.org/10.1002/smj.2014.
[6] G. Hoberg, G. Phillips, Text-based network industries and endogenous product differentiation, J. Political Econ. 124 (2016) 1423–1465, https://doi.org/10.1086/688176.
[7] Y. Wang, L. Kung, W.Y.C. Wang, C.G. Cegielski, An integrated big data analytics-enabled transformation model: application to health care, Inf. Manage. 55 (2018) 64–79, https://doi.org/10.1016/j.im.2017.04.001.
[8] S. Gregor, A.R. Hevner, Positioning and presenting design science research for maximum impact, MIS Q. 37 (2013) 337–355, https://doi.org/10.2753/MIS0742-1222240302.
[9] M.D. Myers, J.R. Venable, A set of ethical principles for design science research in information systems, Inform. Manage. 51 (2014) 801–809, https://doi.org/10.1016/j.im.2014.01.002.
[10] M. Farjoun, Beyond industry boundaries: human expertise, diversification and resource-related industry groups, Organ. Sci. 5 (1994) 185–199.
[11] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (2008) 155–168.
[12] M. Dalziel, A systems-based approach to industry classification, Res. Policy. 36 (2007) 1559–1574.
[13] D. Chong, H. Zhu, Firm Clustering Based on Financial Statements, Social Science Electronic Publishing, 2012.
[14] B. Wernerfelt, The resource-based view of the firm: ten years after, Strat. Manage. J. 16 (1995) 171–174, https://doi.org/10.1002/smj.4250160303.
[15] E. Ahmed Ben, N. Ahlem, G. Faïez, Group extraction from professional social network using a new semi-supervised hierarchical clustering, Knowledge Info. Syst. 40 (2013) 29–47, https://doi.org/10.1007/s10115-013-0634-x.
[16] C. Ge, K.-W. Huang, I.P.L. Png, Engineer/scientist careers: patents, online profiles, and misclassification bias, Strat. Manage. J. 37 (2016) 232–253, https://doi.org/10.1002/smj.2460.
[17] S. Antoine, S. Cécile, G. Hilda, Advanced logistics in Italy: a city network analysis, Tijdschrift Voor Economische En Sociale Geografie. 108 (2017) 753–767, https://doi.org/10.1111/tesg.12215.
[18] Z. Wang, S. Li, G. Zhou, Personal summarization from profile networks, Front. Comput. Sci. 11 (2016) 1085–1097, https://doi.org/10.1007/s11704-016-5088-3.
[19] S. Fortunato, D. Hric, Community detection in networks: a user guide, Phys. Rep. 659 (2016) 1–44, https://doi.org/10.1016/j.physrep.2016.09.002.
[20] M.E.J. Newman, Detecting community structure in networks, Eur. Phys. J. B - Condensed Matt. 38 (2004) 321–330, https://doi.org/10.1140/epjb/e2004-00124-y.
[21] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[22] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (2004) 6, https://doi.org/10.1103/PhysRevE.70.066111 Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics.
[23] A. Medus, G. Acuña, C.O. Dorso, Detection of community structures in networks via global optimization, Phys. A: Stat. Mech. Appl. 358 (2005) 593–604, https://doi.org/10.1016/j.physa.2005.04.022.
[24] S. Sobolevsky, R. Campari, A. Belyi, C. Ratti, General optimization technique for high-quality community detection in complex networks, Phys. Rev. E 90 (2014) 1–8, https://doi.org/10.1103/PhysRevE.90.012811 Statistical, Nonlinear, and Soft Matter Physics.
[25] P. Pons, M. Latapy, Computing communities in large networks using random walks, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).3733, LNCS (2005) 284–293, https://doi.org/10.1007/11569596_31.
[26] C.T. Martin, Rosvall and bergstrom, maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 1118–1123.
[27] S.H. Bae, D. Halperin, J. West, M. Rosvall, B. Howe, Scalable flow-based community detection for large-scale network analysis, Proceedings - IEEE 13th International Conference on Data Mining Workshops (2013) 303–310, https://doi.org/10.1109/ICDMW.2013.138.
[28] S.-H. Bae, B. Howe, GossipMap, Proceedings of the International Conference for High Performance Computing (2015) 1–12, https://doi.org/10.1145/2807591.2807668.
[29] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (2004) 5, https://doi.org/10.1103/PhysRevE.69.066133 Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics.
[30] S. Bhojraj, C.M.C. Lee, D.K. Oler, What's my line? A comparison of industry classification schemes for capital market research, J. Accounting Res. 41 (2003) 745–774, https://doi.org/10.1046/j.1475-679X.2003.00122.x.
[31] D. Lamponi, Is industry classification useful to predict U.S. stock price co-movements? J. Wealth Manage. 17 (2014) 71–77, https://doi.org/10.3905/jwm.2014.

17.1.071.
[32] B.H. Mayhew, R.L. Levinger, Size and the density of interaction in human aggregates, Am. J. Sociol. 82 (1976) 86–110, https://doi.org/10.1086/226271.
[33] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159, https://doi.org/10.2307/2529310.

Xiaoying Xu is an Assistant Professor in the Department of Decision Science, School of Business Administration, South China University of Technology (SCUT). He received his Ph.D. degree in Information Systems from the National University of Singapore (NUS) and his B. Eng. degree in Software Engineering from the South China University of Technology (SCUT). His research interests include text mining, social network analysis, and recommendation. His work has been published in leading journals such as Journal of the Association for Information Science and Technology and Electronic Commerce Research and Applications and has appeared in top conference proceedings such as International Conference on Information Systems.

Hanlin Qian is a Master Candidate at School of Business Administration at the South China University of Technology (SCUT). He received his bachelor's degree in Industrial Engineering from the South China University of Technology (SCUT). His research interests include social network analysis and text mining. His work has been published in the Proceedings of International Conference on Information Resources Management.

Chunmian Ge is a Professor in the Department of Financial Management, School of Business Administration, South China University of Technology (SCUT). He received his Ph.D. degree (2014) in Information Systems from the National University of Singapore (NUS) and his B. Eng. degree (2009) in Computer Science and International Economy and Trade from SCUT. He is a CFA charterholder. His work has appeared or are forthcoming in leading journals such as Strategic Management Journal, Research Policy, IEEE Transactions on Engineering Management, Journal of the Association for Information Science and Technology, and the Proceedings of the International Conference on Information Systems. He serves or has served as Associate Editor for several information systems conferences.

Zhijie Lin is an Associate Professor in the Department of Management Science and Engineering, School of Economics and Management, Tsinghua University. Prior to joining Tsinghua University in 2019, he was an Associate Professor in the Department of Marketing and Electronic Business, School of Business, Nanjing University. He received his Ph.D. degree in Information Systems from the National University of Singapore. His research interests focus on economics of information systems, sharing economy, electronic commerce, and social media. He has published his works in journals such as MIS Quarterly, Information Systems Research, Journal of Management Information Systems, Journal of the Association for Information Systems, Journal of Strategic Information Systems, Research Policy, Decision Support Systems, Information & Management, Electronic Commerce Research and Applications, and in the proceedings of conferences such as International Conference of Information Systems (ICIS) and Annual Meeting of the Academy of Management (AOM). His works have also been covered by many media outlets, including Accenture, Consumer Value Creation, Convince & Convert, Brighton, and LSE Business Review. He is the recipient of the 2011–2015 Young Scholar Innovation Award, 2016 Young Scholar Innovation Award, 2017 Young Scholar Innovation Award, and 2018 Wu Jiapei Award from China Information Economics Society. He serves as a reviewer for NSFC and about 30 journals and conferences, and has served as Associate Editor at ICIS and Track Co-Chair at PACIS.