Information Processing and Management xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Information Processing and Management

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

Using weighted k-means to identify Chinese leading venture capital firms incorporating with centrality measures

Yang Hu[a], Luo Jar-Der[b,*], Fan Ying[c], Zhu Li[d]

[a] School of Information, Central University of Finance and Economics, Beijing, China
[b] Department of Sociology, Tsinghua University, Beijing, China
[c] School of Systems Science, Beijing Normal University, Beijing, China
[d] National School of Development, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Venture capital
Clustering
Ranking
Influential nodes identification
Complex networks
Centrality measures

ABSTRACT

Although identifying leading venture capital firms (VCs) is a meaningful challenge in the analysis of the Chinese investment market, this research topic is rarely mentioned in the relevant literature. Given the co-investment network of VCs, identifying leading VCs is equal to determine influential nodes in the field of complex network analysis. As there are some disadvantages and limitations of using single centrality measures and the multiple criteria decision analysis (MCDA) method to identify leading VCs, this paper incorporates with several different centrality measures of co-investment network of VCs, and then proposes a new approach based on the weighted k-means to rank VCs at both group and individual levels and identify the leading VCs. The proposed approach not only shows alternative groupings based on multiple evaluation criteria, but also ranks them according to their comprehensive score which is the weighted sum of these criteria. Empirical analysis shows the efficiency and practicability of the proposed approach to identify leading Chinese VCs.

1. Introduction

Previous studies have found that the Chinese venture capital market are typically led by leading venture capital firms or investors (leading VCs), who have good opportunities, play the role of main investor(s), set up an investment plan or organize the limited partners (LPs, or 'followers'), and some other VCs often have a tendency to follow a leading VCs (Luo, Rong, Yang, Guo & Zou, 2018). This is the effect of preferential attachment (Barabási, 2005; Powell, Koput, White & Owensmith, 2005) which allows leading VCs to improve their position in the investment market, and thus play important roles in the industry. This is known as so-called co-investment and occurs because it helps share investment risk (Wilson, 1968), builds a word-of-mouth reputation network to hedge against opportunistic behavior (Tykvová, 2007), and creates a better position for competition (Bygrave, 1987; Hochberg, Ljungqvist & Yang, 2007). It is also viewed as a pool of productive resources in which network members can share resources with others to compensate for each other's insufficiencies (Dufour, Nasica & Torre, 2011). For these reasons, identifying leading VCs and collaborating with them is the best strategy to hedge against this high investment uncertainty (Luo, Zhou, Tang & Zhou, 2014; Peters, 2017).

Identifying leading VCs concerns organizing VCs into several groups in order to prioritize them and make sense of their landscape. Despite the importance of identifying leading VCs, this research topic is rarely mentioned in the relevant literature. As we know, this

* Corresponding author.
E-mail address: jdluo@mail.tsinghua.eud.cn (J.-D. Luo).

https://doi.org/10.1016/j.ipm.2019.102083
Received 9 September 2018; Received in revised form 4 July 2019; Accepted 10 July 2019

paper is the first study to identify leading VCs. Although some previous studies have been valuable for the general understanding of VCs, such as through using empirical analysis to find patterns associated with venture performance, venture success, venture growth, or venture value (Batjargal, 2007, 2010, Batjargal et al., 2013; Sievers, Mokwa & Keienburg, 2009), or through using data mining techniques to predict the co-investment relationship (Wang, Zhou, Tang & Luo, 2015) and find triadic closure patterns (Hong, Jie, Lu, Luo & Fu, 2015), they still have not mentioned the task of identifying leading VCs directly. In order to undertake this task, we can refer to various methods mentioned in other fields.

In the Chinese VC industry, leading VCs generally share the three following role characteristics. Firstly, they act as major investors who initiate and organize investment plans, and call for a group of frequent cooperators to follow. Thus, a leading VC and its frequent cooperators form a circle (Luo et al., 2018). Secondly, it is the circle leaders rather than other circle members who play the bridging role connecting various circles. Thirdly, these leading VCs bond together tightly to form an elite clique as the center of whole industry (Gu & Liu, 2019; Useem, 1984). In the following part of this paper, we will use these features to define a leading VC, whose information is collected by Delphi survey.

Given the co-investment network of VCs, identifying leading VCs is equal to determine influential nodes in the field of complex network analysis (Aral & Walker, 2012; Kitsak et al., 2010; Lü, Zhang, Chi & Tao, 2011). With great theoretical and practical significance, determining influential nodes has been widely used in biological, social and technological network analysis(Barrat & Alain, 2008; Pastorsatorras, 2001; Tao, Zhongqian & Binghong, 2006; Vespignani, 2012). The most common method of identifying influential nodes is using centrality measures (Hou, Yao & Liao, 2012; Szolnoki, Xie, Ye & Perc, 2013; Zhong, Gao, Zhang, Shi & Huang, 2014), such as degree centrality, closeness centrality (Sabidussi, 1966), betweenness centrality (Freeman, 1978), eigenvector centrality (Bonacich, 2007), k-core (Kitsak et al., 2010; Perc, 2009), PageRank(Page, 1998), UW-PageRank (Wang, Liu & Wang, 2007), IPRA (Zhong & Lv, 2018), HillTop (Bharat & Mihaila, 2000), TrustRank (Gyöngyi, Garcia-Molina & Pedersen, 2004), and LeaderRank (Bian, Hu & Deng, 2017; Li, Zhou, Lü & Chen, 2013; Lü et al., 2011).

However, choosing which centrality measure to identify influential nodes is a challenging problem. Multiple Criteria Decision Analysis (MCDA), which is defined as an extension of decision theory that covers any decision with multiple objectives (Keeney & Raiffa, 1993), is proposed to identify influential nodes. To overcome the deficiencies of single centrality measure while identifying the leading nodes, Du (2014) used DC, CC, BC in TOPSIS to generate ranks, allowing them to evaluate a node's influence (Du, Gao, Hu, Mahadevan & Deng, 2014; Kuo, 2017). After that, the classic MCDA method: the Analytic Hierarchy Process (AHP) has also been proposed to identify influential nodes (Bian et al., 2017). This is the task of clustering analysis (Meyer & Olteanu, 2013) which is used to automatically divide observations into groups without labeling VCs, so that similar data objects are within one cluster, and dissimilar ones are assigned to different clusters, allowing for the possibility of separating out noise (Assent, 2012; Gullo, Tagarelli & Greco, 2009; Jain, Murty & Flynn, 1999; Ren, Domeniconi, Zhang & Yu, 2017; Thalamuthu, Mukhopadhyay, Zheng & Tseng, 2006; Van der Laan, Pollard & Bryan, 2003). In addition, clustering analysis, the other example, has seen a few applications in the field of MCDA. However, it is used with concepts native to that field, such as the notions of similarity and distance measures (Meyer & Olteanu, 2013).

In this study, the new approach for clustering and ranking Chinese VCs is developed based on the weighted k-means algorithm which can divide alternatives into different groups and estimate evaluation criteria weight. After performing the clustering analysis, the so-called comprehensive score of VCs which is the weighted sum of multi evaluation criteria is produced. Therefore, clusters or VCs can be simply and intuitively ranked according to their score. The originality of the paper comes from the proposed decision model and application of the model for leading VCs identification and ranking in Chinese venture capital market. The rest of this paper is organized as follows: Section 1 summarizes the important research concerning leading VC identification and related studies. Section 2 presents the problem definition and the framework of clustering and ranking leading VCs. Section 3 describe the data and indicators extracted from data. Section 4 proposes the process of implementing clustering and ranking VCs to identify leading VCs. Section 5 presents the empirical studies which illustrate the clustering stability and accuracy of the proposed approach. Section 6 contains the conclusion.

2. Identifying leading VCs

2.1. Problem definition

Suppose we have a set of n VCs denoted by $O = \{o_1, o_2, ..., o_n\}$ and VCs' co-investment relationships represented by a graph $G = (V, E)$ where $V = \{v_1, v_2, ..., v_n\}$ is a set of nodes corresponding to O, and $E \subset V \times V$ is a set of relationships connecting VCs. Notation $e(v_i, v_j) \in E$ (or simply denoted as e_{ij}). The goal is to prioritize the VCs and find the leading group of VCs. According to this definition, the process of identifying leading VCs involves two tasks: 1) ranking VCs and 2) clustering VCs. They are respectively defined as follows.

Definition 1. Ranking VCs. This task first involves obtaining a ranking list $\{rank(o_1), rank(o_2), ..., rank(o_n)\}$ of VCs. $rank(o_i)$ is determined by a MCDA ranking function $f(\sum_{j=1}^{Q} w_j x_j)$, such as TOPSIS (Du et al., 2014; Kuo, 2017), and $X = \{x_1, x_2, ..., x_m\}$ is a set of decision indicators extracted from G, and $w = (w_1, w_2, ..., w_m)$ is a vector of weight corresponding to m decision indicators.

Given the decision indicators and their weight, we can use TOPSIS to rank VCs. The TOPSIS method contains following 5 steps, shown in Algorithm 1.

Definition 2. Clustering VCs. Given the number of clusters K ($\leq n$), this task of identify leading VCs also concerns dividing a set of VCs $O = \{o_1, o_2, ..., o_n\}$ into different groups $C = \{C_1, C_2, ..., C_K\}$ so as to minimize the within-cluster sum of squares (WCSS), such as k-

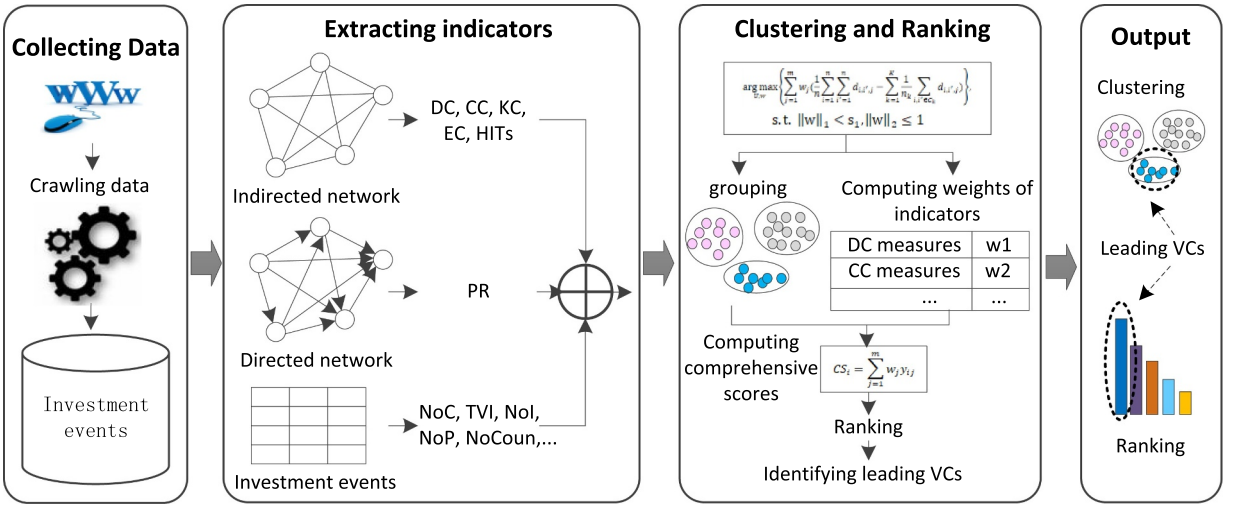

Fig. 1. Framework for identifying leading VCs based on clustering and ranking method.

means, which is $\min_{C} \{\sum_{k=1}^{K} \sum_{o_i \in C_k} d(o_i, u_k)\}$ where u_k is the mean of points in cluster C_k and $d(o_i, u_j)$ is the dissimilarity measurement between o_i and u_k., shown in Algorithm 1.

In practice, we will encounter some issues in applying both algorithms to identify leading VCs. Centrality measures, when used as single criteria, are limited in their ability to interpret nodes' importance. MCDA methods, integrating multiple criteria together to describe a node's position in a network, are more reasonable than single criteria. Even so, besides the criteria of co-investments, criteria related to scale and experience are also important to describe the position of VCs (Bygrave, 1987; Cable & Shane, 1997; Cumming & Dai, 2010). Including such indicators beyond centrality measures will help us improve the accuracy of identifying leading VCs. On the other hand, ranking alternatives based on MCDA methods generally means relying on conflicting criteria. The criteria we applied to identify leading VCs are not independent. Moreover, the weighting of criteria also needs to be outlined in advance. In contrast, the k-means algorithm cannot rank observations and also depends on the number of clusters.

2.2. Framework used for identifying leading VCs

To deal with the aforementioned issues, a framework for identifying leading VCs was designed, shown in Fig. 1.

The frame consists of four steps, as described below:

Step 1: Collecting Data. Collecting investment events from websites and building a database of investment events.

Step 2: Extracting indicators. This is a critical step of the proposed framework. In this step, we construct an undirected investment network that does not investment order and a direct investment network that takes into account investment order, and evaluate the position of every VC in both co-investment networks using several centrality measures. Co-investment is defined as two VCs that have invested in the same company in the same place and same period in this paper. We also use quantitative indicators related to the scale and the experience of VCs to understand the individual investment behaviors of VCs. Indicators related to investment frequency are extracted from investment events as well. More details are described in Section 3.

Step 3. Clustering and ranking. This is a critical step, which contains: clustering, ranking and identifying leading VCs. To eliminate or reduce the impact of indicator correlation, the weighted k-means which penalizes weights of indicators of loss function is utilized to identify leading VCs in this study. The algorithm not only retains the advantages of the classic k-means but also simultaneously estimates indicators' importance. The weighted k-means are used to divide VCs into different groups and estimate indicators' weight. The weighting also illustrates how important the indicator is. An indicator with a larger weighting means its leading VC identification discrimination is higher than others. The weights will be applied to construct the comprehensive score which is the weighted sum of all indicators for every VC. After performing the clustering and ranking analysis, we can divide VCs into different clusters which can be ranked by their average weighted sum of multiple criteria, or sort VCs from highest to lowest scoring, to identify VCs with high-M scores, which should also be leading VCs. More details are described in Section 4.

Step 4: Results. We identify leading VCs based on the clustering results and VCs' ranking.

3. Data and feature extraction

This section introduces in detail our data and feature extraction.

3.1. Data collection

In China, major venture capital databases such as ChinaVenture, Zero2IPO and Venture Capital Research Institute's annual reports release data regarding all public investments and relevant indexes in the venture capital field from 2001 to 2012. We first gather 908 VCs which have 7640 investing events that took place in the Chinese venture capital market from 2001 to 2012. Each event indicates that a VC firm has invested in a company. Investment information includes which company a VC has invested in at what time and which place, and also lists which industry the company belongs to and in which investing period (such as initial stage, expansion stage, seed stage) it is at.

3.2. Indicators of co-investments relationship

A leading VC is typically responsible for post-investment monitoring and due diligence, while syndicate members can help collect and monitor information (Cumming & Dai, 2010). Thus, the position of nodes in a social network is very important for efficient communication during a co-investment. To extract the indicators from observations' relationships G, we need to first define the co-investments of VCs. Let $t = 1, 2, 3. .,T$ be the investing timestamp, we defined indicator I_{ij} as: $I_{ij}(t) = 1$ if VC i and VC j have invested in the same company during the same investment period that the company belongs to and at same place at timestamp t; $I_{ij}(t) = 0$ otherwise.

Definition 3. Undirected co-investment network $G^U = (V^U, E^U)$. $V^U = \{v_1, v_2, ..., v_n\}$ is a set of VCs and $E^U \subset V \times V$ is a set of relationships connecting VCs. For each entry $e_{ij} \in E^U$, $e_{ij} = \sum_{t=1}^{T} I_{ij}(t)$. E^U is a symmetry matrix that means $e_{ij} = e_{ji}$ for $\forall\ i, j = 1, 2, ..., n$.

Let t^i and t^j be the timestamp that VC i and VC j have invested in the same company during the same investment period that the company belongs to and at same place, we defined the indicator as $I_{ij}(t^i \leq t^j) = 1$ if investment of VC i occurs earlier than that of VC j and if the investment occurs at the same time we defined the indicator as $I_{ij}(t^i = t^j) = I_{ji}(t^i = t^j) = 1$.

Definition 4. Directed co-investment network $G^D = (V^D, E^D)$. Difference from Definition 1, $e_{ij} = \sum_{t^i=1}^{T} \sum_{t^j=t^i}^{T} I_{ij}(t^i \leq t^j)$. E^D is an asymmetry matrix that means e_{ij} may not equal to e_{ji} for $i, j = 1, 2, ..., n$. $e_{ij} = e_{ji}$ if and only if VC i and VC j have invested in the same company during the same investment period that the company belongs to and at same place and at same timestamp.

Under Definition 3, we constructed an undirected co-investment network without considering the investing timestamp order. The undirected co-investment network contains 908 nodes and 11,258 ties during the 12 years, including 674 nodes co-investing with others and 234 nodes investing independently. All of them have invested at least on one occasion. For the 674 co-investing nodes, the average number of companies the VC has invested in is 10.42, and for 234 independently investing nodes is 1.86. The 234 nodes can be viewed as isolated points of a network. The most common centrality measures, such as degree centrality (DC), closeness centrality (CC), k-core centrality (KC), eigenvector centrality (EC) and hub scores (HITs), are used to measure each VCs' position in the co-investment network.

(1) Degree centrality (DC). Degree centrality evaluates a node's influence in a complex network (Gao, Ma, Chen, Wang & Xing, 2014; Huang, Fu & Sun, 2015; Newman, 2010). Let $C_d(j)$ denote the degree centrality of node j. $NB_h(j)$ denotes the set of neighbors of node j at a h-hop distance. $C_d(j)$ is therefore defined as: $C_d(j) = |NB_h(j)|$ where $NB_h(j) = \sum_{i=1}^{n} e_{ij}$ is the number of neighbors of node j at the h-hop distance, in most cases, $h = 1$; and if node i connects to j, $e_{ij} = 1$. A high degree centrality indicates a large number of connections between a node and its neighbors.

(2) Closeness centrality (CC). Closeness centrality measures the average length of the shortest paths from one node to other nodes (Gao et al., 2014; Huang et al., 2015; Newman, 2010). The closeness centrality $CC(x)$ of member x is dependent on its geodesic distance, i.e. the shortest paths from member x to all other people in the social network and is calculated as: $CC(x) = \frac{m-1}{\sum_{y \neq x, y \in M} c(x,y)}$ where $c(x, y)$ is a function describing the distance between nodes x and y, m is the number of nodes in a network.

(3) K-core centrality (KC). A k-core (Kitsak et al., 2010; Perc, 2009) of a graph G is a maximal connected subgraph of G in which all vertices have a degree of at least k. Also, it is one of the connected components of the subgraph of G formed by repeatedly deleting all vertices of degree less than k. If a non-empty k-core exists, then G has a degeneracy of at least k, and the degeneracy of G is the largest k for which G has a k-core. A vertex u has coreness c if it belongs to a c-core but not to any $(c + 1)$-core (Newman, 2010).

(4) Eigenvector centrality (EC). Let x be eigenvector of the largest eigenvalue λ of the non-negative adjacency matrix A of the undirected graph $G = (V,\ E)$ where $Ax = \lambda x$, $x_i = u \sum_{j=1}^{n} a_{ij} x_j$ with proportionality factor $u = \frac{1}{\lambda}$ so that x_i is proportional to the sum of similarity scores of all nodes connected to it for $i = 1, 2, ..., n$ (Bonacich, 2007).

(5) Hub scores (HITs). For a web page v in our subset of the web, we use $h(v)$ to denote its hub score and $a(v)$ to denote its authority score. Initially, we set $h(v) = a(v) = 1$ for all nodes v. We also denote by $v \mapsto y$ the existence of a hyperlink from v to y. The core of the iterative algorithm is a pair of updates to the hub and authority scores of all pages given by $h(v) \leftarrow \sum_{v \mapsto y} a(y)$ and $a(v) \leftarrow \sum_{y \mapsto v} h(y)$, which captures the intuitive notions that good hubs point to good authorities and that good authorities are pointed to by good hubs.

In addition, according to Definition 4, we constructed a directed co-investment network that took into consideration the investing timestamp order. Taking into account order of investments, in this paper, PageRank centrality is applied to measure centrality of nodes. PageRank measures (PR) VCs' position in the directed co-investment network and contains 908 nodes and 8696 ties.

(1) PageRank (PR). For a directed network with n nodes and m edges, to make the network strongly connected, we add a new node, called a ground node, and link all others by bidirectional edges. The new network is strongly connected, which is with $n + 1$ nodes and $m + 2n$ edges. Matrix $A = (a_{ij})$ captures the network's wiring diagram. $a_{ij} = 1$ if node i points to node j and it means that user j is a fan of i in social networks. The PR assigns a score to every node, where score implies importance. Mathematically, the PR value of node v_i at this step is: $PR_i(t) = \sum_{j=1}^{N} a_{ji} \frac{PR_j(t-1)}{k_j^{out}}$ where n is the total number of nodes in the network, and k_j^{out} is the out-degree of node j. The above iteration will stop if the PR values of all nodes reach the steady state (Page, 1998).

Because PageRank has remarkable stability properties and it makes a suitable candidate to rank nodes (Mariani, Medo & Zhang, 2015), and some variants of PageRank such as LeaderRank (Linyuan, Zhang, Ho & Tao, 2011) and CiteRank (Jomsri, Sanguansintukul & Choochaiwattana, 2011) have similar performance to PageRank. However, after analyzing, in this study, we did not include these criteria to evaluate the centrality of directed co-investment network.

3.3. Indicators from individual investment behaviors

To make up for limitations of co-investment information from investing events, we also considered various indicators derived from investment events. Especially, most investors prefer to invest in several firms not one, because they want to diversify investment risks. Some VCs have more resources than others to do this. They tend to have a larger scale and greater experience than those who can only invests in smaller amounts. Measures related to scale and experience are therefore important to describe the position of VCs (Bygrave, 1987; Cable & Shane, 1997; Cumming & Dai, 2010). For instance, the number of companies the VC has invested in (NoC) indicates whether or not a VC has sufficient assets and resources to invest (Wright & Lockett, 2003). If a VC has a high NoC, it should have the power and influence to accumulate resources to invest in larger deals. Therefore this indicator is a good variable to distinguish leading VCs from followers. Similarly, the total number of investments (TNI), the number of industries a VC has invested in (NoI), the number of periods a VC has invested in (NoP), the number of countries a VC has invested in (NoCoun), the number of provinces a VC has invested in (NoPR), the number of stages: initial stage (NoSI), expansion stage (NoSE) and seed stage (NoSS) a VC has invested in all measure or describe the scale and experience of VCs.

3.4. Summary of indicators

In this paper, six indicators related to VCs' positions in co-investment networks and nine indicators related to the investing scale and experience of VCs are used to identify leading VCs (see Fig. 2). The summary statistics of 15 indicators are calculated based on 908 Chinese VCs are presented in Table 1, including mean, standard deviation (sd), median, min, max, skew and kurtosis.

4. Implementation of the clustering and ranking algorithm

4.1. Clustering and estimating weight of indicators

(1) The weighted k-means. The weighted k-means algorithm is employed to divide VCs and also to determine which indicators are important to the clustering. This algorithm retains the performance advantages of k-means whilst overcoming its shortcomings and dealing with correlated indicators data. In addition, this algorithm does not need to know the data distribution before

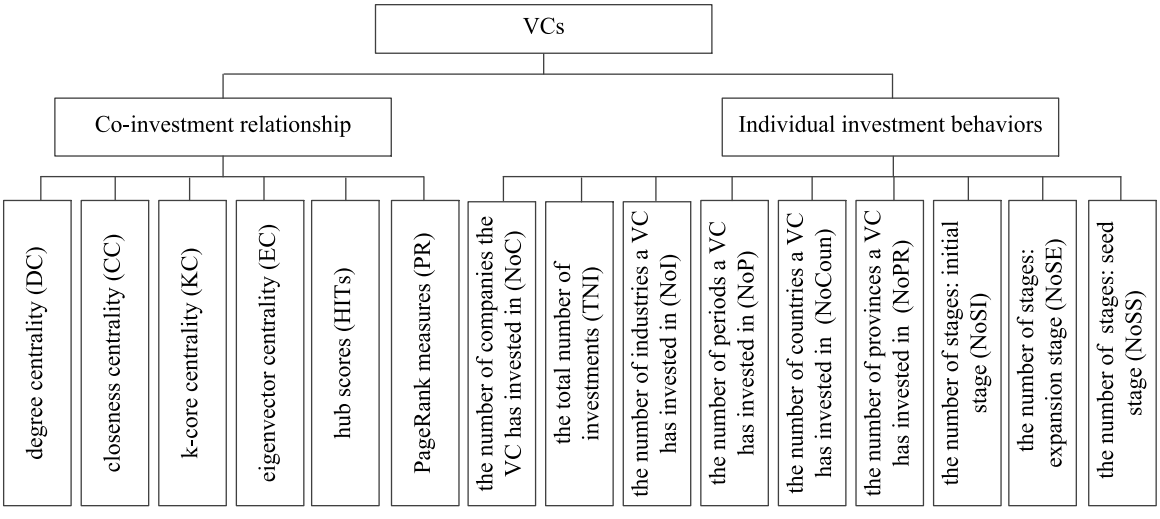

Fig. 2. Indicators of individual investment and networks applied to identify leading VCs.

Table 1
Summary of the 15 indicators of 908 Chinese VCs.

Indicators	Mean	Sd	Median	Min	Max	Skew	Kurtosis
KC	13.6938	24.4092	4.0000	0.0000	110.0000	2.7267	6.974
DC	24.7974	65.7298	4.0000	0.0000	728.0000	5.1971	34.1776
CC	0.0000	0.0000	0.0000	0.0000	0.0000	−0.7042	−1.5058
EC	0.0293	0.0974	0.0008	0.0000	1.0000	5.4063	34.5246
HITs	0.0293	0.0974	0.0008	0.0000	1.0000	5.4063	34.5246
PR	0.0011	0.0018	0.0005	0.0002	0.0254	5.5886	47.8437
TVI	8.2159	22.4582	2.0000	1.0000	321.0000	6.9184	66.5654
NoC	6.2919	16.2479	2.0000	1.0000	275.0000	8.0702	99.1598
NoI	5.1344	10.2183	2.0000	1.0000	106.0000	4.5382	25.8302
NoP	1.3678	0.507	1.0000	1.0000	3.0000	0.8242	−0.6898
NoCoun	1.0573	0.2718	1.0000	1.0000	4.0000	5.5000	35.2598
NoPR	2.2907	2.8357	1.0000	1.0000	26.0000	3.7242	17.286
NoSE	4.2654	11.9099	1.0000	0.0000	221.0000	9.3962	134.6277
NoSS	0.0297	0.2207	0.0000	0.0000	2.0000	7.887	63.6042
NoSI	2.3436	6.3739	1.0000	0.0000	87.0000	6.2678	55.3375

clustering in contrast to others (Reihanian, Feizi-Derakhshi & Aghdasi, 2017; Yang, Mcauley & Leskovec, 2014). Let w_j be the weight along feature j, we define dissimilarity measure $d_{i,i',j} = (x_{ij} - x_{i'j})^2$ between observation i and observation i' along indicator j (for $i = 1, 2, ...,n$; $k = 1, 2, ...,K$; $j = 1, 2, ...,m$). The weighted between-cluster sum of squares (Witten & Tibshirani, 2010) is therefore defined as

$$\arg\max_{U,w}\left\{\sum_{j=1}^{m} w_j\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i',j}\right)\right\}, \tag{1}$$

$$s.\ t.\ \ w_1 < s_1,\ \ w_2 \le 1$$

where $w_1 = \sum_{j=1}^{m}|w_j| < s_1$ is the lasso penalty (Assent, 2012; Johnstone & Titterington, 2009; Ma & Huang, 2008) to control the scale of features. w_2 is the l-2 norm penalty and $w_2 = \sqrt{\sum_{j=1}^{p} w_j^2}$.

In formula (1), $\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i',j}$ is the average sum of the square and is a measure of variability of all observations, and $\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i',j}$ is the within-cluster sum of squares and is a measure of the variability of the observations within each cluster. Let the different between the average sum of the square and the within-cluster sum of squares be $a_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i',j}$ for $j = 1, 2, ...,p$. Our goal is to maximize (1), or minimize the within-cluster sum of squares $\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i',j}$ to allocate the similar observations within a group. The weight is calculated by

$$w_j = \frac{S(a_j, \lambda_1)}{\sqrt{\sum_{j=1}^{m} S^2(a_j, \lambda_1)}} \tag{2}$$

where $S(x, c)$ is the soft threshold function $S(x, c) = sign(x)(|x| - c)_+$. According to (2), if indicator j is good at distinguishing observations from others, w_j should have a large value.

(1) Determining the number of clusters. According to leader member exchange theory (Graen, 1976; Graen & Cashman, 1975; Liden, Sparrowe & Wayne, 1997), relationships between supervisors and their subordinates foster different levels of interpersonal trust through tangible and intangible social exchanges (Dienesch & Liden, 1986). Leading VCs are those who lead their followers in the Chinese venture investment market. We therefore decide the number of clusters, which was derived from social studies. The first famous one is Dunbar's number. It is a cognitive limit to the number of people with whom one can maintain stable social relationships—relationships in which an individual knows who each person is and how each person relates to every other person. According to this theory, we set the number of clusters K from 4 to 8 according to the number of VCs.
(2) Choosing parameters. The optimal tuning parameters λ_1 and λ_2 are obtained when they make the gap statistic (Tibshirani, Walther & Hastie, 2000, 2001) maximal.

4.2. Ranking VCs

The nature of TOPSIS (Du et al., 2014; Kuo, 2017) concerns defining the positive idea solution and the negative idea solution, calculating the separation measures using the Euclidean distance, and then calculating the ranking index to rank observations in algorithm. In our study, the larger the value of VC's indicators means that leading VCs have more co-investing opportunities, a larger

operating scale and richer experience, the more chance the VC is a leader in the venture capital industry. That is similar to determining the defining the positive idea solution. We therefore simplify TOPSIS to compute a comprehensive score which is the weighted sum of all observed values along all indicators, defined as

$$CS_i = \sum_{j=1}^{m} w_j y_{ij}, \tag{3}$$

where $y_{ij} = \frac{z_{ij}}{\sqrt{\sum_{i=1}^{n} z_{ij}^2}}$ and w_j is the weight estimated by the weighted k-means. Similar to rank index in TOPSIS, the comprehensive score is therefore applied to rank individual VCs or the group of VCs.

4.3. Identifying leading VCs

With a comprehensive score, leading VCs can be determined by the elbow method (Kodinariya & Makwana, 2013) or identifying VCs with high M (Top-M) comprehensive score. This is meaningful because leading VCs have larger scale, richer experience and more co-investing opportunities than others from empirical results (Bygrave, 1987; Cable & Shane, 1997; Cumming & Dai, 2010; Gao et al., 2014; Huang et al., 2015; Newman, 2010). Therefore, they will have significant and large observation values along each evaluation criterion. Given the weight of criteria, their comprehensive score has a higher probability of being larger than others. On the basis of this assumption, leading VCs can be determined by looking at a high M comprehensive score after sorting VCs in descending order by their comprehensive scores. This approach is called Clustering and Ranking Individual Leading VCs (CRILVC). In addition, because of the similarity of VCs, leading VCs are divided into clusters. According to their average comprehensive score, we can rank VCs in clusters or groups level, plot mean-plots of them and find the point of elbow in this plot. The average comprehensive score of leading clusters in the left side of the point of elbow is clearly larger than others in the right side of the point of elbow. This is another way to identify leading VCs, called Clustering and Ranking Group Leading VCs (CRGLVC).

5. Empirical results

5.1. Accuracy analysis

To evaluate the accuracy of the proposed method, we obtain the reference label of venture capital firms was obtained using the Delphi inquiry method (Linstone & Turoff, 1975) whose results are obtained from interviews in April 2013. Before the interviews, we computed the k-core centrality of each VC (Gu, Luo, and Liu 2018), and made a list of 908 VCs according to their k-core centrality. Because it is very difficult to find informants who are very familiar to 908 VCs, we only interviewed four experts, who are well known investors or researchers and familiar with the Chinese venture capital industry. We asked them to choose a subset of “most influential” VCs as defined by the three features stated above from the list without limiting the size of their choice, and giving the reason for their choice. They checked “yes” or “no” for all listed VCs. Among the four informants, one is the leader of a VC research institute in the Chinese central government, one is the CEO of a large foreign investor, and the other two are CEOs of large domestic VCs. Unfortunately, two of them were not familiar with some of foreign VCs, and thus they marked “unknown” as their response for those investors.

Interviewers first explain the definition of the leading VCs. As suggested above, leading roles in the Chinese VC industry are the main investors, who initiate, plan and organize almost all runs of their investments, and call for some followers, often frequent cooperators, to follow. Those followers gather together around the focal VC and form a solid group for more investment opportunities (Luo et al., 2018). The informants selected a subset of “leading” VCs from the list according to their experience. Finally, we selected the VCs without any response of “no” from the informants, which resulted in a total of 42 VCs being selected. The summary statistics of the 15 indicators calculated based on the reference labels are presented in Table 2, including mean, standard deviation (sd), median, min, max, skew and kurtosis. We therefore take 42 VCs as the reference labels of leading investors in this section, and the others are taken to be as the remaining VCs.

(1) **Comparing CRGLVC with other clustering algorithms.** We compare the performance of CRGLVC with other clustering algorithms to illustrate the stability of the approach when using different clustering numbers. Similar to the weighted k-means, others, including: k-means, ensemble k-means, partition around medoids (PAM) clustering algorithm and ensemble PAM (Gullo et al., 2009; Ren et al., 2017; Van der Laan et al., 2003), are also used to identify leading VCs. All clustering results are conducted from the same process shown in Fig. 1. Given the number of clusters K, all these clustering algorithms are applied to divide Chinese venture capital firms into K clusters. Three experimental scenarios are $K = 4$, $K = 6$ and $K = 8$ respectively. After clustering, we compute the average comprehensive score for each cluster, and sort clusters in descending order according to their average score. All these results are presented in Fig. 3, which contains four charts, the first three are the mean-plot for the three scenarios. Each error-bar in these figures represent the mean and standard deviation of each clusters’ comprehensive score. In addition, the elbow point is also marked in the first three mean-plots. In the last chart (Fig. 3), each bar represents the number of leading VCs based on a certain clustering method for each scenario.

In Fig. 3, only observations in the first two clusters with the highest average comprehensive score are regarded as leading VCs in scenario 1 ($K = 4$) according to the elbow method. Similarly, observations in the first three clusters and observations in the first five

Table 2
Summary of the 15 indicators based on the leading VCs and others.

Ref	Indicators	Mean	Sd	Median	Min	Max	Skew	Kurtosis
The leading VCs	KC	86.429	27.549	98.000	12.000	110.000	−0.814	−0.533
	DC	247.095	154.133	213.000	16.000	728.000	0.996	0.679
	CC	0.000	0.000	0.000	0.000	0.000	−0.079	−0.060
	EC	0.324	0.253	0.271	0.002	1.000	0.819	−0.175
	HITs	0.324	0.253	0.271	0.002	1.000	0.819	−0.175
	PR	0.007	0.005	0.006	0.001	0.025	1.785	4.612
	TVI	76.714	64.580	59.000	9.000	321.000	1.747	3.490
	NoC	53.619	48.655	40.000	7.000	275.000	2.509	8.066
	NoI	35.762	23.482	28.500	4.000	106.000	0.893	0.078
	NoP	2.143	0.354	2.000	2.000	3.000	1.969	1.924
	NoCoun	1.714	0.805	2.000	1.000	4.000	0.810	−0.222
	NoPR	9.810	5.641	8.000	1.000	26.000	0.847	0.068
	NoSE	36.690	38.250	26.000	5.000	221.000	2.850	10.397
	NoSS	0.214	0.565	0.000	0.000	2.000	2.405	4.431
	NoSI	21.333	17.491	17.500	1.000	87.000	1.663	3.203
Others	KC	10.166	17.873	4.000	0.000	110.000	3.384	13.016
	DC	14.016	29.805	4.000	0.000	294.000	4.446	25.151
	CC	0.000	0.000	0.000	0.000	0.000	−0.628	−1.607
	EC	0.015	0.050	0.001	0.000	0.593	6.387	50.651
	HITs	0.015	0.050	0.001	0.000	0.593	6.387	50.651
	PR	0.001	0.001	0.001	0.000	0.008	3.058	12.728
	TVI	4.894	9.608	2.000	1.000	95.000	4.674	26.494
	NoC	3.997	7.112	1.000	1.000	66.000	4.412	23.578
	NoI	3.649	5.966	1.000	1.000	55.000	4.171	21.199
	NoP	1.330	0.483	1.000	1.000	3.000	0.874	−0.844
	NoCoun	1.025	0.157	1.000	1.000	2.000	6.022	34.303
	NoPR	1.926	2.011	1.000	1.000	20.000	3.877	20.206
	NoSE	2.693	5.085	1.000	0.000	53.000	4.882	30.454
	NoSS	0.021	0.185	0.000	0.000	2.000	9.500	93.065
	NoSI	1.423	3.121	0.000	0.000	28.000	4.627	27.003

clusters with the highest score are regarded as the leading VCs in scenario 2 ($K = 6$) and scenario 3 ($K = 8$) respectively. The corresponding number of leading VCs of CRGLVC, k-means, ensemble k-means, PAM and ensemble PAM are (28,102,114,41,38) in scenario 1, (30,100,102,90,31) in scenario 2, and (34,87,100,106,43) in scenario 3. The results indicate that CRGLVC is stable as the diversity ensemble clustering PAM. CRGLVC tends to obtain the nearly same number of leading VCs in different scenarios. It also identifies fewer high comprehensive scoring VCs compared to other methods. The reason why it performs stably and accurately is it considers the different importance of indicators and also reduces the influence of indicator correlation.

(1) **Comparing** CRILVC **and** CRGLVC **with single centrality measures and TOPSIS.** To prove the accuracy and robustness of the proposed approach, results of VCs identification based on CRILVC and CRGLVC are compared with the single centrality measures such as DC, CC, BC, EC, SLC, PR and LR, and TOPSIS. Because centrality measures and TOPSIS are unable to divide VCs into different groups, we not only show the results of CRGLVC, which identifies leading VCs at a group level, but also present results using CRILVC, which distinguishes leading VCs at an individual level. To do this, we sort the overall score of individual VCs in descending order, so we can identify VCs with high M scores. Similarly, VCs with high M scores are taken as leaders based on single centrality measures and TOPSIS. Here, $M = 25, 50, 80$ in our study. Because the score of some VCs may be the same, M many more or less different from what we have given. Various evaluation criteria have been introduced to evaluate the performance of the different approaches.

Let TP be the number of leading VCs correctly identified as leading VCs; FP be the number of leading VCs incorrectly identified as leading VCs; TN be the number of other VCs correctly identified as other VCs; and FN be the number of leading VCs incorrectly identified as other VCs. We use recall - the fraction of leading VCs that have been retrieved over the total amount of leading VCs, and precision – the fraction of leading VCs among the retrieved leading VCs. Both of them are defined as $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ and $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ (Raghavan, Bollmann & Jung, 1989). Recall is also known as sensitivity. It can be viewed as the probability that the leading VCs are identified by the proposed approach. In addition, some other indices are used to evaluate the performance of the approaches, described as following: (1) The **Purity** (Manning, Raghavan & Schutze, 2008) is a measure of the extent to which clusters contain a single class. A purity score of 1 is possible by putting each VC in its own cluster that means the larger purity the method has, the performance of the algorithm is better; (2) The **Jaccard** index (Hamers, 1989) is used to quantify the similarity between the identified leading VCs and the reference labels of leading VCs. The Jaccard index takes on a value between 0 and 1. An index of 1 means that the leading VCs are totally correctly identified, and an index of 0 indicates that the results are totally incorrect. (3) The normalized mutual information (Rong, Nguyen & Jaatun, 2013; Estévez, Tesmer, Perez & Zurada, 2009) derived from mutual information, is used here to measure how much information is shared between a clustering and a reference classification.

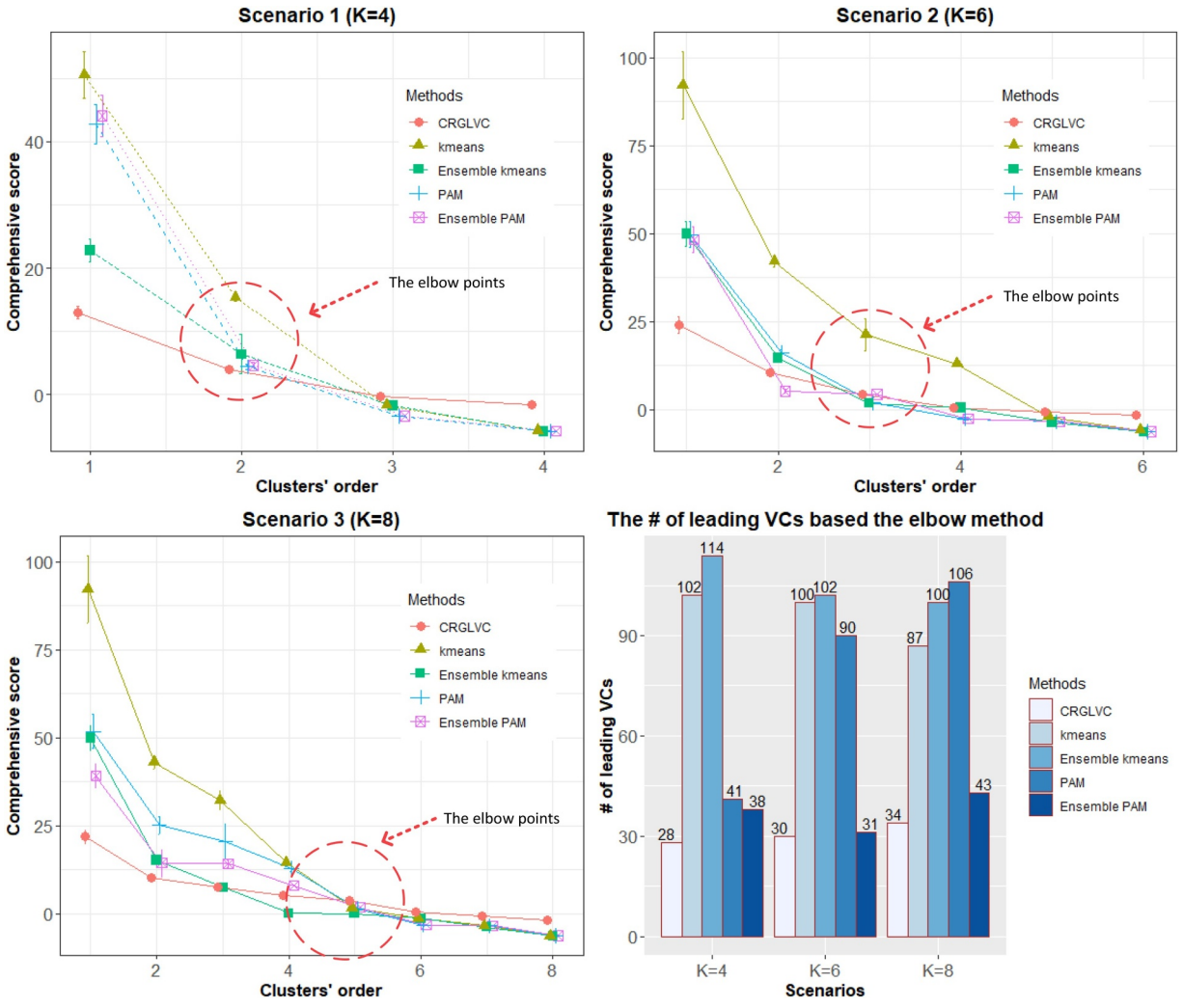

Fig. 3. Mean plots of four clustering algorithms and the number of leading VCs based on elbow method.

The average value and standard deviation of all evaluation criteria for each approach are summarized based on 100 datasets randomly sampled from our original VC dataset given the number of clusters (K = 5) and the sampling ratio (0.7, 0.8, 0.9). **Ref** denotes the number of the reference labels of leading VCs that is obtained by inquire in the results. All these results are shown in Table 3–5.

Tables 3–5 show that CRGLVC can automatically identify leading VCs when taking the first two clusters as leaders based on the elbow method. Taking the 42 leading VCs as a reference, 31.46(15.27), 33.67(14.09), 36.82(15.05) leaders are identified with respect to the sample ratio = 0.7, 0.8, 0.9. In contrast to CRGLVC, given top-M = 25, 50 and 80, there are 20.04(1.59), 26.5(2.28), 28.37(2.62) leading VCs, 21.23 (1.11), 29.01 (1.84), 31.79 (2.31) leading VCs, 22.05 (0.78), 31.33 (1.48), 35.3 (1.62) with respect to the sample ratio = 0.7, 0.8, 0.9 correctly identified by CRILVC. Obviously, the performance of CRILVC is similar to CRGLVC. The only difference is whether the number of leading VCs is given automatically or not – in the former it is given, whereas in the latter it is not. Although reducing sample size will influence the performance of CRILVC, it still performs better than others, indicating it does not depend on a few cases. Moreover, the reason why single criteria perform poorly compared to CRILVC is they only focus on one kind of structural information for VCs. Similar to CRILVC, some results are obtained based on single centrality measures and TOPSIS. TOPSIS only performs better than the proposed approach in a few analysis settings. This indicates using the comprehensive score rank and identify the leading VCs is similar to that of TOPSIS, but that the comprehensive score is easier to calculate and interpret.

In addition, it is worth noting that DC and PR perform better than other signal criteria, but not as well as CRILVC and TOPSIS. DC is used to measure the local position of nodes and PR is diffusion-based centrality measures which evaluate a node's position in a directed network. As mentioned, leading VCs' excellent reputation, maneuverable resources and accessible information, they will have more chances to invest first and many followers co-invest with them so as to obtain good opportunities, learn investment technique, and sharing their reputation (Barabási, 2005; Powell et al., 2005). Hence, PR accurately identifies leading VCs to some extent.

Table 3
Leading VC identification based on the proposed method, single centrality measures and TOPSIS (Sampling ratio = 0.7, the average number of reference leading VCs = 29.19 ± 2.72).

Methods	top-M	Ref	Precision	Recall	Purity	Jaccard	NMI
CRGLVC	31.46 (15.27)*	20.22 (4.33)	0.71 (0.16)	0.69 (0.14)	0.97 (0.01)	0.93 (0.03)	0.46 (0.09)
CRILVC	25 (0.00)	20.04 (1.59)	0.8 (0.06)	0.69 (0.05)	0.98 (0.00)	0.95 (0.01)	0.53 (0.06)
	50 (0.00)	26.5 (2.28)	0.53 (0.05)	0.91 (0.04)	0.96 (0.00)	0.91 (0.01)	0.46 (0.04)
	80 (0.00)	28.37 (2.62)	0.35 (0.03)	0.97 (0.02)	0.95 (0.00)	0.84 (0.01)	0.34 (0.03)
TOPSIS	25 (0.00)	19.52 (1.76)	0.78 (0.07)	0.67 (0.05)	0.98 (0.00)	0.95 (0.01)	0.5 (0.06)
	50 (0.00)	25.17 (2.23)	0.5 (0.04)	0.86 (0.04)	0.96 (0.00)	0.91 (0.01)	0.41 (0.04)
	80.02 (0.14)	27.6 (2.65)	0.34 (0.03)	0.95 (0.03)	0.95 (0.00)	0.83 (0.01)	0.31 (0.03)
DC	25.29 (0.61)	19.16 (1.59)	0.76 (0.06)	0.66 (0.06)	0.97 (0.00)	0.95 (0.01)	0.48 (0.06)
	50.62 (0.92)	26.02 (2.42)	0.51 (0.05)	0.89 (0.03)	0.96 (0.00)	0.91 (0.01)	0.43 (0.04)
	82.48 (2.31)	28.24 (2.82)	0.34 (0.04)	0.97 (0.02)	0.95 (0.00)	0.83 (0.01)	0.32 (0.04)
CC	25.12 (0.33)	15.65 (2.19)	0.62 (0.09)	0.54 (0.07)	0.96 (0.01)	0.93 (0.01)	0.33 (0.08)
	50.25 (0.54)	24.68 (3)	0.49 (0.06)	0.85 (0.08)	0.96 (0.00)	0.9 (0.01)	0.39 (0.07)
	80.48 (0.78)	27.08 (2.7)	0.34 (0.03)	0.93 (0.03)	0.95 (0.00)	0.83 (0.01)	0.3 (0.04)
KC	27.05 (2.33)	17.22 (1.85)	0.64 (0.06)	0.59 (0.07)	0.97 (0.01)	0.93 (0.01)	0.37 (0.06)
	52.01 (1.97)	24.94 (2.51)	0.48 (0.05)	0.86 (0.05)	0.96 (0.00)	0.9 (0.01)	0.39 (0.05)
	83.17 (3.03)	28.17 (2.8)	0.34 (0.04)	0.96 (0.03)	0.95 (0.00)	0.83 (0.01)	0.32 (0.04)
EC	25 (0.00)	16.53 (1.83)	0.66 (0.07)	0.57 (0.07)	0.97 (0.01)	0.93 (0.01)	0.37 (0.07)
	50 (0.00)	23.04 (2.23)	0.46 (0.04)	0.79 (0.05)	0.95 (0.00)	0.89 (0.01)	0.34 (0.04)
	80.01 (0.1)	26.89 (2.85)	0.34 (0.04)	0.92 (0.04)	0.95 (0.00)	0.83 (0.01)	0.29 (0.04)
HITs	25 (0.00)	16.53 (1.83)	0.66 (0.07)	0.57 (0.07)	0.97 (0.01)	0.93 (0.01)	0.37 (0.07)
	50 (0.00)	23.04 (2.23)	0.46 (0.04)	0.79 (0.05)	0.95 (0.00)	0.89 (0.01)	0.34 (0.04)
	80.07 (0.36)	26.89 (2.85)	0.34 (0.04)	0.92 (0.04)	0.95 (0.00)	0.83 (0.01)	0.29 (0.04)
PR	25 (0.00)	18.4 (1.78)	0.74 (0.07)	0.63 (0.05)	0.97 (0.00)	0.94 (0.01)	0.45 (0.06)
	50 (0.00)	25.44 (2.25)	0.51 (0.04)	0.87 (0.04)	0.96 (0.00)	0.91 (0.01)	0.42 (0.04)
	80 (0.00)	27.02 (2.65)	0.34 (0.03)	0.93 (0.03)	0.95 (0.00)	0.83 (0.01)	0.3 (0.03)

* This value is calculated by CRGLVC automatically.

Table 4
Leading VC identification based on the proposed method, single centrality measures and TOPSIS (Sampling ratio = 0.8, the average number of reference leading VCs = 33.83 ± 2.53).

Methods	top-M	Ref	Precision	Recall	Purity	Jaccard	NMI
CRGLVC	33.67 (14.09)*	23.08 (3.94)	0.74 (0.13)	0.68 (0.12)	0.97 (0.01)	0.94 (0.02)	0.47 (0.07)
CRILVC	25 (0.00)	21.23 (1.11)	0.85 (0.04)	0.63 (0.04)	0.98 (0.00)	0.95 (0.01)	0.52 (0.04)
	50 (0.00)	29.01 (1.84)	0.58 (0.04)	0.86 (0.04)	0.96 (0.00)	0.93 (0.01)	0.47 (0.04)
	80 (0.00)	31.79 (2.31)	0.4 (0.03)	0.94 (0.02)	0.95 (0.00)	0.86 (0.01)	0.36 (0.03)
TOPSIS	25 (0.00)	21.41 (1.3)	0.86 (0.05)	0.63 (0.04)	0.98 (0.00)	0.95 (0.01)	0.52 (0.05)
	50 (0.00)	28.04 (1.97)	0.56 (0.04)	0.83 (0.03)	0.96 (0.00)	0.92 (0.01)	0.43 (0.03)
	80 (0.00)	31.54 (2.29)	0.39 (0.03)	0.93 (0.02)	0.95 (0.00)	0.86 (0.01)	0.35 (0.02)
DC	25.17 (0.4)	20.49 (1.24)	0.81 (0.05)	0.61 (0.04)	0.98 (0.00)	0.95 (0.01)	0.48 (0.05)
	50.52 (0.73)	29.43 (2.22)	0.58 (0.04)	0.87 (0.03)	0.97 (0.00)	0.93 (0.01)	0.48 (0.04)
	81.26 (1.33)	32.7 (2.41)	0.4 (0.03)	0.97 (0.02)	0.95 (0.00)	0.86 (0.01)	0.38 (0.03)
CC	25.12 (0.36)	16.31 (1.75)	0.65 (0.07)	0.48 (0.05)	0.96 (0.00)	0.93 (0.01)	0.32 (0.06)
	50.19 (0.39)	26.64 (2.56)	0.53 (0.05)	0.79 (0.07)	0.96 (0.01)	0.91 (0.01)	0.39 (0.07)
	80.43 (0.7)	31 (2.49)	0.39 (0.03)	0.92 (0.02)	0.95 (0.00)	0.86 (0.01)	0.33 (0.03)
KC	27.02 (2.17)	17.83 (1.85)	0.66 (0.05)	0.53 (0.05)	0.97 (0.00)	0.93 (0.01)	0.34 (0.05)
	51.74 (2.23)	26.79 (1.98)	0.52 (0.04)	0.79 (0.05)	0.96 (0.00)	0.91 (0.01)	0.38 (0.04)
	82.61 (2.8)	32.35 (2.46)	0.39 (0.03)	0.96 (0.03)	0.95 (0.00)	0.86 (0.01)	0.36 (0.03)
EC	25 (0.00)	17.12 (1.52)	0.68 (0.06)	0.51 (0.05)	0.97 (0.00)	0.93 (0.01)	0.35 (0.06)
	50 (0.00)	25.14 (1.87)	0.5 (0.04)	0.74 (0.03)	0.96 (0.00)	0.91 (0.01)	0.35 (0.03)
	80 (0.00)	30.47 (2.35)	0.38 (0.03)	0.9 (0.03)	0.95 (0.00)	0.85 (0.01)	0.32 (0.03)
HITs	25.01 (0.1)	17.12 (1.52)	0.68 (0.06)	0.51 (0.05)	0.97 (0.00)	0.93 (0.01)	0.35 (0.06)
	50 (0.00)	25.14 (1.87)	0.5 (0.04)	0.74 (0.03)	0.96 (0.00)	0.91 (0.01)	0.35 (0.03)
	80 (0.00)	30.47 (2.35)	0.38 (0.03)	0.9 (0.03)	0.95 (0.00)	0.85 (0.01)	0.32 (0.03)
PR	25 (0.00)	19.8 (1.29)	0.79 (0.05)	0.59 (0.04)	0.97 (0.00)	0.95 (0.01)	0.45 (0.04)
	50 (0.00)	28.7 (2.05)	0.57 (0.04)	0.85 (0.03)	0.96 (0.00)	0.92 (0.01)	0.46 (0.04)
	80 (0.00)	30.98 (2.41)	0.39 (0.03)	0.92 (0.02)	0.95 (0.00)	0.86 (0.01)	0.34 (0.03)

* This value is calculated by CRGLVC automatically.

5.2. Influence of indicators

We also analyzed the performance of CRILVC, CRGLVC and TOPIS by using various groups of indicators. For example, only using individual characteristic indicators (NoC, TNI, NoI, NoP, NoCoun, NoPR, NoSI, NoSE and NoSS), or only using co-investment

Table 5
Leading VC identification based on the proposed method, single centrality measures and TOPSIS (Sampling ratio = 0.9, the average number of reference leading VCs = 37.98 ± 1.78).

Methods	top-M	Ref	Precision	Recall	Purity	Jaccard	NMI
CRGLVC	36.82 (15.05)*	25.83 (4.33)	0.75 (0.12)	0.68 (0.11)	0.97 (0.01)	0.94 (0.02)	0.48 (0.05)
CRILVC	25 (0.00)	22.05 (0.78)	0.88 (0.03)	0.58 (0.02)	0.98 (0.00)	0.95 (0.00)	0.5 (0.03)
	50 (0.00)	31.33 (1.48)	0.63 (0.03)	0.83 (0.02)	0.97 (0.00)	0.94 (0.00)	0.48 (0.03)
	80 (0.00)	35.3 (1.62)	0.44 (0.02)	0.93 (0.01)	0.95 (0.00)	0.88 (0.00)	0.39 (0.02)
TOPSIS	25 (0.00)	22.36 (0.76)	0.89 (0.03)	0.59 (0.02)	0.98 (0.00)	0.95 (0.00)	0.52 (0.03)
	50 (0.00)	30.56 (1.34)	0.61 (0.03)	0.8 (0.02)	0.97 (0.00)	0.93 (0.00)	0.46 (0.02)
	80 (0.00)	34.97 (1.57)	0.44 (0.02)	0.92 (0.02)	0.95 (0.00)	0.88 (0.00)	0.38 (0.02)
DC	25.22 (0.44)	21.01 (0.93)	0.83 (0.03)	0.55 (0.03)	0.97 (0.00)	0.95 (0.01)	0.45 (0.04)
	50.64 (0.85)	32.79 (1.52)	0.65 (0.03)	0.86 (0.02)	0.97 (0.00)	0.94 (0.00)	0.52 (0.03)
	80.87 (0.94)	36.38 (1.72)	0.45 (0.02)	0.96 (0.02)	0.95 (0.00)	0.88 (0.00)	0.42 (0.02)
CC	25.15 (0.36)	17.06 (1.35)	0.68 (0.06)	0.45 (0.04)	0.96 (0.00)	0.93 (0.01)	0.31 (0.04)
	50.29 (0.57)	27.44 (2.11)	0.55 (0.04)	0.72 (0.05)	0.96 (0.00)	0.92 (0.01)	0.37 (0.05)
	80.38 (0.75)	34.6 (1.81)	0.43 (0.02)	0.91 (0.02)	0.95 (0.00)	0.88 (0.00)	0.37 (0.02)
KC	26.69 (2.28)	18.16 (1.64)	0.68 (0.05)	0.48 (0.04)	0.97 (0.00)	0.93 (0.01)	0.33 (0.04)
	51.85 (1.99)	28.18 (1.59)	0.54 (0.03)	0.74 (0.03)	0.96 (0.00)	0.92 (0.01)	0.38 (0.03)
	82.25 (2.16)	35.78 (1.78)	0.44 (0.03)	0.94 (0.02)	0.95 (0.00)	0.88 (0.01)	0.39 (0.03)
EC	25 (0.00)	17.54 (1.26)	0.7 (0.05)	0.46 (0.04)	0.97 (0.00)	0.93 (0.01)	0.33 (0.04)
	50 (0.00)	27.11 (1.27)	0.54 (0.03)	0.71 (0.03)	0.96 (0.00)	0.92 (0.00)	0.36 (0.02)
	80 (0.00)	32.86 (1.72)	0.41 (0.02)	0.87 (0.03)	0.95 (0.00)	0.87 (0.00)	0.33 (0.03)
HITs	25 (0.00)	17.54 (1.26)	0.7 (0.05)	0.46 (0.04)	0.97 (0.00)	0.93 (0.01)	0.33 (0.04)
	50 (0.00)	27.11 (1.27)	0.54 (0.03)	0.71 (0.03)	0.96 (0.00)	0.92 (0.00)	0.36 (0.02)
	80 (0.00)	32.86 (1.72)	0.41 (0.02)	0.87 (0.03)	0.95 (0.00)	0.87 (0.00)	0.33 (0.03)
PR	25 (0.00)	20.91 (1.01)	0.84 (0.04)	0.55 (0.02)	0.97 (0.00)	0.95 (0.00)	0.45 (0.03)
	50 (0.00)	30.76 (1.27)	0.62 (0.03)	0.81 (0.03)	0.97 (0.00)	0.93 (0.00)	0.47 (0.03)
	80.03 (0.17)	34.84 (1.67)	0.44 (0.02)	0.92 (0.02)	0.95 (0.00)	0.88 (0.00)	0.38 (0.02)

* This value is calculated by CRGLVC automatically.

indicators (DC, CC, KC, EC, HITs and PR), or using all indicators to identify leading VCs. Given $K = 5$, clustering results for all samples (including 908 samples and 42 leading VCs) are summarized in a similar way to previous sections, shown in Table 6. In addition, the average weight of indicators and it standard deviation are calculated by using the weighted k-means analyze full data repeated 100 times, presented in Table 7.

The clustering results using all indicators do well in Table 6. Other clustering results, whether using individual characteristics indicators or only using co-investment indicators, do not perform as well as results using all indicators. It reveals that both individual and co-investment indicators are vital for identifying leading VCs. Table 7 also indicates that both individual investment behaviors and co-investment relationship play the main role to identify leading VCs. Most of these indicators have large weight besides NoSS.

5.3. Result of identifying leading VCs

Finally, given the number of clusters ($K = 5$), we use both CRILVC and CRGLVC to analyze the original VC dataset and repeated the process 100 times. We then calculated the average comprehensive score (**Score**) and the standard deviation (**sd**) of the Top-50 VCs - which are obtained by CRILVC and the frequency (**Freq.**) of the leading VCs as identified by CRGLVC, shown in Table 8. In contrast to the labels of interviewers, the reference labels (Sagiroglu & Sinanc, 2013) and the Chinese name of VCs was also included in the table.

Table 6
Only using indicators of individual characteristic indicators, or only using centrality measures, or using all indicators to identify leading VCs based on the proposed method (the number of VCs = 908, the number of reference leading VCs = 42).

Indicators	Methods	Identification	ref	Precision	Recall	Purity	Jaccard	NMI
All indicators	CRGLVC	37.55 (11.46)	27.79 (3.8)	0.76 (0.09)	0.66 (0.09)	0.97 (0.01)	0.95 (0.01)	0.48 (0.04)
	CRILVC	22.32 (0.47)	25 (0.00)	0.89 (0.02)	0.53 (0.01)	0.98 (0.00)	0.95 (0.00)	0.47 (0.02)
		33.32 (0.47)	50 (0.00)	0.67 (0.01)	0.79 (0.01)	0.97 (0.00)	0.94 (0.00)	0.49 (0.01)
		39 (0.00)	80 (0.00)	0.49 (0.00)	0.93 (0.00)	0.95 (0.00)	0.9 (0.00)	0.43 (0.00)
Individual indicators	CRGLVC	18.96 (1.72)	11.96 (1.72)	0.63 (0.03)	0.28 (0.04)	0.96 (0.00)	0.92 (0.00)	0.20 (0.03)
	CRILVC	21 (0.00)	25 (0.00)	0.84 (0.00)	0.5 (0.00)	0.97 (0.00)	0.94 (0.00)	0.42 (0.00)
		28 (0.00)	50 (0.00)	0.56 (0.00)	0.67 (0.00)	0.96 (0.00)	0.92 (0.00)	0.35 (0.00)
		36 (0.00)	80 (0.00)	0.45 (0.00)	0.86 (0.00)	0.95 (0.00)	0.89 (0.00)	0.36 (0.00)
Co-investment indicators	CRGLVC	32 (0.00)	22 (0.00)	0.69 (0.00)	0.52 (0.00)	0.97 (0.00)	0.93 (0.00)	0.36 (0.00)
	CRILVC	25 (0.00)	19 (0.00)	0.76 (0.00)	0.45 (0.00)	0.97 (0.00)	0.94 (0.00)	0.35 (0.00)
		50 (0.00)	28 (0.00)	0.56 (0.00)	0.67 (0.00)	0.96 (0.00)	0.92 (0.00)	0.35 (0.00)
		80 (0.00)	30 (0.00)	0.38 (0.00)	0.71 (0.00)	0.95 (0.00)	0.86 (0.00)	0.25 (0.00)

Table 7
The weight of indicators calculated by the weight k-means (the number of sample size is 908, the number of reference leading VCs = 42).

Indicators	Weight	Indicators	Weight	Indicators	Weight
TVI	0.2896 (0.0188)	NoPR	0.2342 (0.0037)	DC	0.3015 (0.0114)
NoC	0.277 (0.0217)	NoSE	0.2511 (0.0178)	CC	0.2727 (0.1528)
NoI	0.2763 (0.0015)	NoSS	0.0931 (0.1376)	EC	0.2495 (0.0143)
NoP	0.2157 (0.1078)	NoSI	0.2654 (0.0235)	HITs	0.2495 (0.0143)
NoCoun	0.145 (0.0122)	KC	0.2676 (0.0216)	PR	0.2845 (0.0081)

Table 8
Average comprehensive score and frequency of the leading VCs calculated based on proposed approach (the number of sample size is 908, the number of reference leading VCs = 42).

No	Name of VC	Ref	Score (sd)	Freq.	No.	Name of VC	Ref	Score (sd)	Freq.
1	SHENZHEN CAPITAL	1	29.25 (1.29)	100	26	NEA	1	9.63 (0.32)	100
2	IDG	1	27.58 (1.14)	100	27	WALDEN	1	9.49 (0.31)	100
3	SEQUOIA	1	20.09 (0.86)	100	28	COWIN	1	9.27 (0.34)	97
4	LEGEND	1	19.9 (0.76)	100	29	CDH	0	8.15 (0.28)	72
5	NORTHERN LIGHT	1	15.43 (0.53)	100	30	CDF	1	7.79 (0.95)	49
6	FORTUNE	1	14.94 (0.58)	97	31	WI HARPER	0	7.67 (0.25)	72
7	QIMING	1	14.88 (0.49)	100	32	SHSTVC	1	6.65 (0.21)	25
8	MATRIX PARTNERS	1	11.87 (0.79)	99	33	QCOM	1	6.64 (0.49)	49
9	SIG CHINA	0	11.86 (0.45)	100	34	LEAGUER	0	6.48 (0.25)	22
10	SAIF	1	11.5 (0.47)	97	35	LIGHTSPEED	1	6.41 (0.19)	25
11	JAFCO ASIA	1	11.49 (0.39)	100	36	CEYUAN	1	6.23 (0.23)	25
12	iD TECH VENTURES	1	11.33 (0.37)	100	37	DFJ DRAGON	1	6.21 (0.22)	25
13	INTEL	1	11.26 (0.38)	100	38	ZKZS	0	5.95 (0.23)	22
14	DETONG	1	11.01 (0.36)	97	39	REDPOINT	1	5.83 (0.21)	25
15	GREEN PINE	1	10.86 (0.46)	97	40	ORIENTAL FORTUNE	0	5.76 (0.23)	22
16	GGV	1	10.71 (0.33)	100	41	C-VC	1	5.73 (0.21)	22
17	SBCVC	1	10.63 (0.32)	97	42	VERTEX VC	0	5.69 (0.2)	25
18	MORNINGSID	1	10.29 (0.94)	99	43	TSINGHUA VC	0	5.68 (0.25)	22
19	GSR	1	10.23 (0.37)	100	44	TTGG	0	5.28 (0.24)	22
20	ORIZA	1	10.19 (0.43)	97	45	GOBIVC	0	5.22 (0.2)	22
21	JIANGSU GOVTOR	1	10.02 (0.4)	97	46	TIANTU	0	4.75 (0.21)	5
22	KPCB	1	9.95 (0.32)	100	47	CMCAPITAL	0	4.7 (0.18)	5
23	ZERO2IPO	0	9.92 (0.38)	100	48	TSINGCAPITAL	0	4.69 (0.14)	5
24	NEW MARGIN	0	9.74 (0.31)	97	49	FUKIN	0	4.6 (0.18)	8
25	DCM	1	9.71 (0.37)	100	50	BLUERUN	1	4.48 (0.43)	32

Algorithm 1
TOPSIS.

Step 1: Normalize decision matrix. The normalized value is calculated as $y_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}$ for $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$.

Step 2: Compute the weighted normalized decision matrix. The weighted normalized is calculated as $y'_{ij} = w_j y_{ij}$ for $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$.

Step 3: Determine the positive idea solution (PIS) and the negative idea solution (NIS). PIS $= \{y_1^{'+}, y_2^{'+}, ..., y_m^{'+}\} = \{\max_{\forall i} y'_{ij} | j \in J_1, \min_{\forall i} y'_{ij} | j \in J_2\}$ and PIS $= \{y_1^{'-}, y_2^{'-}, ..., y_m^{'-}\} = \{\max_{\forall i} y'_{ij} | j \in J_1, \min_{\forall i} y'_{ij} | j \in J_2\}$. J_1 are the set of benefit criteria and J_2 are the set of cost criteria.

Step 4: Calculate the separation measures using the Euclidean distance. The separation of an alternative o_i from the PIS is given as $D_i^+ = \sqrt{\sum_{j=1}^{m} (y'_{ij} - y_1^{'+})^2}$ for $i = 1, 2, ..., n$. Similarly the separation of an alternative o_i from the NIS is given as $D_i^- = \sqrt{\sum_{j=1}^{m} (y'_{ij} - y_1^{'-})^2}$ for $i = 1, 2, ..., n$.

Step 5: Calculate the ranking index. The ranking index is defined as $rank(o_i) = \frac{D_i^-}{D_i^+ + D_i^-}$ where $0 \le rank(o_i) \le 1$ for $i = 1, 2, ..., n$.

Algorithm 2
K-means clustering.

Repeat until convergence: Step 1: Assign each observation to the cluster whose mean has the least squared Euclidean distance, $C_k = \{o_i: d(o_i, u_k) \le d(o_i, u_j) \ \ \forall j, 1 \le j \le k\}$ where each o_i is assigned to exactly one C_k (for $k = 1, 2, ..., K$).

Step 2: Calculate the new means (centroids) of the observations in the new clusters, $u_k = \frac{1}{|C_k|} \sum_{o_i \in C_k} o_i$.

Algorithm 3
Process of clustering and ranking individual leading VCs (CRILVC).

Input: observations, number of clusters, top-M
Output: clustering results and top-M leading VCs
Process:
1 Extract Indicators.
(1a) Calculate indicators of individual investment behaviors such as: NoC, TNI, NoI, NoP, NoCoun, NoPR, NoSI, NoSE and NoSS.
(1b) Construct a directed network and an undirected network of VCs according to Definition 3 and Definition 4, and then compute centrality measures such as: DC, CC, KC, EC, HS and PR.
2 Clustering. Group VCs by the weighted k-means in formula (1) and calculate the weight of indicators by formula (2) where the number of clusters is given according to leader member exchange theory and choose tuning parameters by maximizing the gap statistics.
3 Ranking.
(31) Compute comprehensive scores of each VC according to formula (3) where the weight of indicators is estimated in step 2.
(3b) Sort individual VCs in descending order according to their comprehensive score.
4 Identifying individual leading VCs. Given the parameter of M, take the first M with high comprehensive scores as leading VCs.

Algorithm 4
Process of clustering and ranking group leading VCs (CRGLVC).

Input: observations, number of clusters
Output: clustering results
Process :
1 Extract Indicators.
(1a) Calculate indicators of individual investment behaviors such as: NoC, TNI, NoI, NoP, NoCoun, NoPR, NoSI, NoSE and NoSS.
(1b) Construct a directed network and an undirected network of VCs according Definition 3 and Definition 4, and then compute centrality measures such as: DC, CC, KC, EC, HS and PR.
2 Clustering. Group VCs by the weighted k-means in formula (1) and calculate the weight of indicators by formula (2) where the number of clusters is given according to leader member exchange theory and choose tuning parameters by maximizing the gap statistics.
3 Ranking.
(3a) Compute comprehensive scores of each VC according to formula (3) where the weight of indicators is estimated in step 2.
(3b) Calculate the average comprehensive score of VCs in each cluster.
(3c) Sort clusters in descending order according to their average comprehensive score.
4 Identifying the group of leading VCs.
(4a) Plot mean-plots of clusters' average comprehensive score.
(4b) Find the point of elbow in the plot
(4c) Choose the group of leading VCs whose average comprehensive score is larger than or equal to the point of elbow.

The clustering results show that almost of all leading VCs have a higher comprehensive score than others and some followers are also identified as leading VCs because of their high comprehensive score. The leading VCs always are grouped into the set of leading VCs, so that their frequency is higher than that of others. These results also indicate that the proposed approach can identify the most leading VCs correctly through the 15 indicators extracted from co-investment and individual investment behaviors.

6. Discussion and conclusion

This paper first proposed a framework for guiding how to identify leading VCs in both group and individual levels and identify the leading VCs. The proposed approach consists of four steps: extracting indicators related to both co-investment relationships and individual investment behavior from historical data of investment events, dividing VCs into several groups and estimating the weight of indicators, constructing the comprehensive score similar to TOPSIS to rank VCs, and finally identifying leading VCs based on the elbow method or the top-M. The empirical analysis, in Tables 3–5, shows that the performance that the proposed approach achieves with respect to TOPSIS and some centrality measures is accurate and practicable for the identification of leading Chinese VCs. Although other methods sometimes operate in a similar fashion, they display worse performance than the approach used here in almost all scenarios. In addition, results in Table 6 and 7 also reveal that the accuracy of identifying leading VCs using both co-investment indicators and individual characteristics is better than that only using one kind of indicators, and results in Table 8 indicates that the proposed approach is good at identifying Chinese leading VCs.

However, for future research more improvements do need to be made.

Although the indicators related to social relationship, scale and experience extracted from co-investment relationships and individual behaviors can explain the difference between leading VCs and followers to some extent, they still can not cover all information of VCs in the market. Including indicators beyond what we have mentioned may help us improve the accuracy of identifying leading VCs.

Because it is a difficult to find informants to evaluate such a number of alternatives, such as 908 VCs, the reference labels applied to assess the proposed approach obtained through interviewing only 4 informants may be not convincing enough. This is still a challenging issue in the field of social science.

This paper can be said to only be touching the surface in terms of identifying the leading VCs as defined by the three features stated above. We may require to do several runs of Delphi surveys and data mining to improve the accuracy of our predictions. In

each run, more subject evaluation can be integrated into the study, and more objective predictors can be obtained from data mining. In short, the insufficiencies of the survey in this study can be corrected through continued work and analysis.

None the less, the proposed method is worth considering, especially as is helpful for social scientists to understand leading VCs based on the results of analyzing historical data of investment events. Informants can therefore make decisions and select a set of alternatives more easily by using the results from the proposed method. Furthermore, social scientists can also use the results from this study as a reference to check or make improvements to relevant questionnaires. Similarly, data mining techniques will also be advanced with improvements made in the questionnaire.

Finally, this paper does not consider other methods related to the identification of leading nodes and the measuring of centrality. It only refers to classic machine learning algorithms. More findings that are interesting could be found by using other methods to analyze Chinese venture capital firms in future studies, such as: why some leading VCs are grouped into being called followers while those who are followers are identified as the leading VCs.

Acknowledgments

We thank the two anonymous reviewers whose comments and suggestions helped greatly improve the clarity of this manuscript. This work is supported by a grant from the Young Scientists Fund of the National Natural Science Foundation of China “Studies on Regularized Statistical Methods for Analyzing High Dimensional Big Data”, Project number(71701223). This article also gets the support of Chinese National Science Foundation project “Social Network in Big Data Analysis: A Case of Investment Network”, Project number (71372053) and the support of National Statistical Science Foundation of China “Studies on the Network Structure-Based Dimensionality Reduction Method and Application for High-dimensional Data”, Project number (2018LZ08).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2019.102083.

References

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science, 337*(6092), 337–341.
Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(4), 340–350.
Barabási, A. (2005). Taming complexity. *Nature Physics, 1*(2), 68–70.
Barrat, & Alain (2008). *Dynamical processes on complex networks.* Cambridge University Press.
Batjargal, B. (2007). Comparative social capital: Networks of entrepreneurs and venture capitalists in China and Russia. *Management and Organization Review, 3*(3), 397–419.
Batjargal, B. (2010). The effects of network's structural holes: Polycentric institutions, product portfolio, and new venture growth in China and Russia. *Strategic Entrepreneurship Journal, 4*(2), 146–163.
Batjargal, B., Hitt, M. A., Tsui, A. S., Arregle, J.-L., Webb, J. W., & Miller, T. L. (2013). Institutional polycentrism, entrepreneurs' social networks, and new venture growth. *Academy of Management Journal, 56*(4), 1024–1049.
Bharat, K., & Mihaila, G. (2000). Hilltop: A search engine based on expert documents. *Proceedings of the 9th international WWW conference poster. 10*.
Bian, T., Hu, J., & Deng, Y. (2017). Identifying influential nodes in complex networks based on AHP. *Physica A Statistical Mechanics & Its Applications, 479*(4), 422–436.
Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks, 29*(4), 555–564.
Bygrave, W. D. (1987). Syndicated investments by venture capital firms: A networking perspective. *Journal of Business Venturing, 2*(2), 139–154.
Cable, D. M., & Shane, S. (1997). A prisoner's dilemma approach to entrepreneur-venture capitalist relationships. *Academy of Management Review, 22*(1), 142–176.
Cumming, D., & Dai, N. (2010). Local bias in venture capital investments. *Journal of Empirical Finance, 17*(3), 362–380.
Dienesch, R. M., & Liden, R. C. (1986). Leader-member exchange model of leadership: A critique and further development. *Academy of Management Review, 11*(3), 618–634.
Du, Y., Gao, C., Hu, Y., Mahadevan, S., & Deng, Y. (2014). A new method of identifying influential nodes in complex networks based on TOPSIS. *Physica A Statistical Mechanics & Its Applications, 399*(4), 57–69.
Dufour, D., Nasica, E., & Torre, D. (2011). Optimal syndication choices in venture capital investment: Understanding the role of skills and funds providers. .
Estévez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks, 20*(2), 189–201.
Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239.
Gao, S., Ma, J., Chen, Z., Wang, G., & Xing, C. (2014). Ranking the spreading ability of nodes in complex networks based on local structure. *Physica A Statistical Mechanics & Its Applications, 403*(6), 130–147.
Graen, G. (1976). Role-making processes of leadership development. *Handbook of Industrial and Organizational Psychology,* 1201–1245.
Graen, G., & Cashman, J. F. (1975). A role-making model of leadership in formal organizations: A developmental approach. *Leadership Frontiers, 143*, 165–196.
Gu, W., & Liu, J. (2019). Exploring small-world network with an elite-clique: Bringing embeddedness theory into the dynamic evolution of a venture capital network. *Social Networks, 57*, 70–81.
Gullo, F., Tagarelli, A., & Greco, S. (2009). Diversity-based weighting schemes for clustering ensembles. *Proceedings of the 2009 SIAM international conference on data mining* (pp. 437–448). Society for Industrial and Applied Mathematics.
Gyöngyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating web spam with trustrank. *Proceedings of the thirtieth international conference on very large data bases. 30. Proceedings of the thirtieth international conference on very large data bases* (pp. 576–587). VLDB Endowment.
Hamers, L. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management, 25*(3), 315–318.
Hochberg, Y. V., Ljungqvist, A., & Yang, L. U. (2007). Whom you know matters: Venture capital networks and investment performance. *Journal of Finance, 62*(1), 251–301.
Hong, H., Jie, T., Lu, L., Luo, J. D., & Fu, X. (2015). Triadic closure pattern analysis and prediction in social networks. *IEEE Transactions on Knowledge & Data Engineering, 27*(12), 3374–3389.
Hou, B., Yao, Y., & Liao, D. (2012). Identifying all-around nodes for spreading dynamics in complex networks. *Physica A Statistical Mechanics & Its Applications, 391*(15), 4012–4017.
Huang, C. Y., Fu, Y. H., & Sun, C. T. (2015). Identify influential social network spreaders. *2014 IEEE International conference on data mining workshop* (pp. 562–568). IEEE.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *Acm Computing Surveys, 31*(3), 264–323.
Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A, 367*(1906), 4237–4253.
Jomsri, P., Sanguansintukul, S., & Choochaiwattana, W. (2011). CiteRank: Combination similarity and static ranking with research paper searching. *International Journal of Internet Technology & Secured Transactions, 3*(2), 161–177.
Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives : Preferences and value tradeoffs.* Cambridge University Press.
Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Eugene Stanley, H., et al. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*(11), 888–893.
Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal, 1*(6), 90–95.
Kuo, T. (2017). A modified TOPSIS with a different ranking index. *European Journal of Operational Research, 260*(1), 152–160.
Lü, L., Zhang, Y. C., Chi, H. Y., & Tao, Z. (2011). Leaders in social networks, the delicious case. *PloS one, 6*(6), e21202.
Li, Q., Zhou, T., Lü, L., & Chen, D. (2013). Identifying influential spreaders by weighted leaderrank. *Physica A Statistical Mechanics & Its Applications, 404*(24), 47–55.
Liden, R. C., Sparrowe, R. T., & Wayne, S. J. (1997). Leader-member exchange theory: The past and potential for the future. *Research in Personnel and Human Resources Management, 15*, 47–120.
Linstone, H. A., & Turoff, M. (1975). *The delphi method.* MA: Addison-Wesley Reading.
Linyuan, L., Zhang, Y. C., Ho, Y. C., & Tao, Z. (2011). Leaders in Social Networks, the DeliciousCase. *PloS one, 6*(6), e21202.
Luo, J. D., Zhou, L., Tang, J., & Zhou, Y. (2014). Why do chinese venture capitals invest jointly? An analysis of complex investment network. *Academy of Management Annual Meeting Proceedings, 2014*(1), 13807.
Luo, J. D., Rong, K., Yang, K., Guo, R., & Zou, Y. Q. (2018). Syndication through social embeddedness: A comparison of foreign, private and state-owned venture capital (VC) firms. *Asia Pacific Journal of Management, 36*(2), 1–29.
Ma, S., & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Brief Bioinform, 9*(5), 392–403.
Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval.* Cambridge university press.
Mariani, M. S., Medo, M., & Zhang, Y. C. (2015). Ranking nodes in growing networks: When pagerank fails. *Scientific Reports, 5*(16181), 1–10.
Meyer, P., & Olteanu, A. L. (2013). Formalizing and solving the problem of clustering in MCDA. *European Journal of Operational Research, 227*(3), 494–502.
Newman, M. (2010). *Networks: An introduction.* Oxford University Press.
Page, L. (1998). The pagerank citation ranking : Bringing order to the web. *Stanford Digital Libraries Working Paper, 9*(1), 1–14.
Pastorsatorras, R. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters, 86*(14), 3200–3203.
Perc, M. (2009). Evolution of cooperation on scale-free networks subject to error and attack. *New Journal of Physics, 11*(3), 033027.
Peters, R. H. (2017). *Volatility and venture capital.* Retrieved from http://repository.upenn.edu/fnce_papers/31.
Powell, W. W., Koput, K. W., White, D. R., & Owensmith, J. (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology, 110*(4), 1132–1205.
Raghavan, V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS), 7*(3), 205–229.
Reihanian, A., Feizi-Derakhshi, M. R., & Aghdasi, H. S. (2017). Community detection in social networks with node attributes based on multi-objective biogeography based optimization. *Engineering Applications of Artificial Intelligence, 62*, 51–67.
Ren, Y., Domeniconi, C., Zhang, G., & Yu, G. (2017). Weighted-object ensemble clustering: Methods and analysis. *Knowledge and Information Systems, 51*(2), 661–689.
Rong, C., Nguyen, S. T., & Jaatun, M. G. (2013). Beyond lightning: A survey on security challenges in cloud computing. *Computers & Electrical Engineering, 39*(1), 47–54.
Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika, 31*(4), 581–603.
Sagiroglu, S., & Sinanc, D. (2013). "Big data: A review. *2013 International conference on collaboration technologies and systems (CTS)* (pp. 42–47). IEEE.
Sievers, S., Mokwa, C. F., & Keienburg, G. (2009). The relevance of financial versus non-financial information for the valuation of venture capital-backed firms. *European Accounting Review, 22*(3), 467–511.
Szolnoki, A., Xie, N. G., Ye, Y., & Perc, M. (2013). Evolution of emotions on networks leads to the evolution of cooperation in social dilemmas. *Physical Review E Statistical Nonlinear & Soft Matter Physics, 87*(4), 042805.
Tao, Z., Zhongqian, F., & Binghong, W. (2006). Epidemic dynamics on complex networks. *Progress in Natural Science:MAterials International, 16*(5), 452–457.
Thalamuthu, A., Mukhopadhyay, I., Zheng, X., & Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics, 22*(19), 2405–2412.
Tibshirani, R., Walther, G., & Hastie, T. (2000). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, 63*(2), 411–423.
Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423.
Tykvová, T. (2007). Who chooses whom? Syndication, skills and reputation. *Review of Financial Economics, 16*(1), 5–28.
Useem, E. L. (1984). Education and high-technology industry: The case of Silicon Valley. *Economics of Education Review, 3*(3), 215–221.
Van der Laan, M., Pollard, K., & Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation, 73*(8), 575–584.
Vespignani, A. (2012). Modelling dynamical processes in complex socio-technical systems. *Nature Physics, 8*(1), 32–39.
Wang, J., Liu, J., & Wang, C. (2007). Keyword extraction based on pagerank. *Pacific-Asia conference on knowledge discovery and data mining4426*. *Pacific-Asia conference on knowledge discovery and data mining* (pp. 857–864). Springer.
Wang, Z., Zhou, Y, Tang, J., & Luo, J. D. (2015). The prediction of venture capital co-investment based on structural balance theory. *IEEE Transactions on Knowledge and Data Engineering, 28*(2), 537–550.
Wilson, R. (1968). The theory of syndicates. *Econometrica : Journal of the Econometric Society, 36*(1), 119–132.
Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association, 105*(490), 713–726.
Wright, M., & Lockett, A. (2003). The structure and management of alliances: Syndication in the venture capital industry. *Journal of Management Studies, 40*(8), 2073–2102.
Yang, J., Mcauley, J., & Leskovec, J. (2014). "Community detection in networks with node attributes. *2013 IEEE 13th international conference on data mining* (pp. 1151–1156). IEEE.
Zhong, L., & Lv, F. (2018). An improved pagerank for identifying the influential nodes based on resource allocation in directed networks. *2017 14th international computer conference on wavelet active media technology and information* (pp. 42–45). IEEE.
Zhong, L., Gao, C., Zhang, Z., Shi, N., & Huang, J. (2014). Identifying influential nodes in complex networks: A multiple attributes fusion method. *International conference on active media technology* (pp. 11–22). Springer.