Association for Information Systems
AIS Electronic Library (AISeL)

WHICEB 2020 Proceedings | Wuhan International Conference on e-Business

Summer 7-5-2020

Identifying Chinese Leading Venture Capital Firms Based on Graph Convolutional Neural Networks

Caijiang Cheng
School of Information, Central University of Finance and Economics, Beijing, 102206, China, 1319305263@qq.com

Hu Yang
School of Information, Central University of Finance and Economics, Beijing, 102206, China

Xin Jin
School of Information, Central University of Finance and Economics, Beijing, 102206, China

Xiaoyi Tang
School of Information, Central University of Finance and Economics, Beijing, 102206, China

Identifying Chinese Leading Venture Capital Firms Based on Graph Convolutional Neural Networks

Caijiang Cheng[1*], *Hu Yang*[2], *Xin Jin*[3], *Xiaoyi Tang*[4]
[1,2,3,4]School of Information, Central University of Finance and Economics, Beijing, 102206, China

Abstract: It is a meaningful challenge to identify leading venture capital firms (VCs) in the analysis of the Chinese investment market. Identifying leading VCs is equal to determine influential nodes in the field of complex network analysis. Many studies have applied centrality measures to determine influence nodes. However, only a few studies have explored more efficient and flexible ways to accomplish this task. In this work, we propose a new approach which using graph convolutional neural networks to identify influential nodes in the network, so as to determine leading VCs. We build an undirected graph based on co-investment of VCs, then learn a VCs Graph Convolutional Neural Network (vcGCNN) for nodes classification. Our vcGCNN is labeled with '1' and '0' for 'is leading VCs' and 'is not leading VCs'. The experiment results on VCs dataset demonstrate that vcGCNN outperforms multiple centrality measures and some typical spectral-based GNN methods for leading venture capital firms identification.

Keywords: venture capital firms, graph convolutional neural networks, influential nodes identification

1. INTRODUCTION

Previous research found that Chinese venture capital market is usually dominated by leading Venture Capital Firms (VCs), who are more likely to obtain good investment opportunities and play the role of main investors, that is, to formulate investment plans or organize investment partners, while other VCs are tending to follow leading VCs [1]. On the one hand, the effect of preferential attachment [2] allows main investors to improve status in the market which is beneficial to main investors. On the other hand, co-investment among VCs is helpful to share investment risks [3], establish reputation network against opportunistic behaviors [4], improve the competitiveness in the industry [5], share resources among members and learn from each other [6]. Therefore, to deal with high investment risk, it is the best strategy to identify leading venture capital firms and cooperate with them in venture capital market [7].

Identifying leading venture capital firms is equal to determine influential nodes while the co-investment network of VCs is given [8]. In the field of complex network analysis, centrality measures are the most common used methods in identifying important nodes, however, centrality measures usually only evaluate the importance of nodes from one single aspect, and different network structures may have various centrality measure methods, it's a challenging job to decide a best one. At the same time, the centrality measures ignore the characteristics of the node itself, but these characteristics are also important features to identify the influence of nodes. Since centrality measures in complex network analysis only concern about structure between nodes, Graph Neural Networks (GNNs) which integrate the network topology information and the feature of the nodes, has been used for node classification, and reached higher accuracy in the classification of network nodes.

Due to the rapid changes in Chinese venture capital market, new investment companies have joined the venture capital team. In the absence of labels, it is practical to identify new leading VCs through semi-supervised learning algorithms. Graph convolutional networks (GCN) is a robust learning algorithm that can efficiently identify the categories of unlabeled nodes by partial labeled nodes [9].

[1] Corresponding author. Email: 1319305263@qq.com (Caijiang Cheng)

In this work, we propose a new approach to identify leading venture capital firms basing on graph convolutional neural networks. We build a dataset of investment events collected form website, then construct a large graph of co-investment between VCs from the dataset, which contains a set of nodes as VCs and a set of edges as co-investment relationship, also, frequency of co-investment is set as weight on edges. Then we model the graph with a Graph Convolutional Network (GCN), which captures information of graph efficiently. In this way, we have turned a leading venture capital firms identification problem into a node classification problem. Nodes are labeled with '1' and '0' for 'is leading VCs' and 'is not leading VCs'. Result shows that our method can achieve strong classification performances with limited labeled nodes. To summarize, our contributions are as follows:

- We propose a new approach for leading VCs identification. As far as we know, this is the first study to model an investment events dataset as an undirected graph and construct model with graph convolutional neural networks to learn nodes classification.
- Results on investment events dataset demonstrate that our method outperforms most common used methods of centrality measures. With a small proportion of labeled nodes, our method also learns predictive nodes efficiently.

The rest of this paper is organized as follows: Section 2 summarizes related works about centrality measures and graph neural networks. Section 3 proposes the method of identifying leading VCs. Section 4 presents the empirical settings and results of proposed method comparing with baseline methods, which illustrate the accuracy of the proposed method. At last, section 5 presents the conclusion and future work expected.

2. RELATED WORK

2.1 Traditional Centrality Measures

The most common used methods of identifying influential nodes are centrality measures, which have been widely used in social， biological and technological network analysis [10,11,12]. Centrality measures includes degree centrality [13], betweenness centrality [14] closeness centrality [13], eigenvector centrality [15], K-core [16] , PageRank [17], etc. Unlike these methods, our method not only uses the structure information between nodes, but also makes good use of nodes' characteristics information, which is important to determine influential nodes in the network.

2.2 Graph Neural Networks

As graph convolutions are more efficient and convenient to composite with other neural networks, convolutional graph neural networks (ConvGNNs) has been rapidly growing in recent years, which falls into two categories [18], spectral-based ConvGNNs and spatial-based ConvGNNs. The core of spectral-based ConvGNNs is how to define the convolutional operation, for example: ChebNet [19] approximates the convolutional filter by Chebyshev polynomials, GCN [9] makes several approximations and simplifications to reduce computation complexity, CayleyNet [20] further applies Cayley polynomials to capture narrow frequency bands, etc. In addition, spatial-based approaches inherit ideas from recurrent graph neural networks by information aggregation to define graph convolutions.

3. METHOD

3.1 Graph Convolutional Networks (GCN)

The pioneering work of GCN has presented a simplified graph convolutional neural networks model, which achieved excellent nodes classification results on a number of benchmark citation datasets [9].

First of all, let's consider an undirected graph $G = (V, \mathcal{E})$, where $V = (v_1, v_2, \ldots, v_N)$ is a set of nodes and $\mathcal{E}$ is the edge set. We introduce the adjacency matrix $A \in \mathbb{R}^{N\times N}$ of graph G which is nonnegative, and the degree matrix $D = diag(d_1, d_2, \ldots, d_N)$ of adjacency matrix, where $d_i = \sum_j a_{ij}$. So that the graph Laplacian is defined as $L = D - A$. We introduce the feature matrix $X \in \mathbb{R}^{N\times M}$ where N is the number of nodes and M is the dimension of feature vectors, so that $x_v \in \mathbb{R}^M$ is the feature vector with M dimension for node v. GCN can capture information about one-hop neighbors with one layer of convolution, and larger scale of neighbors while multiple GCN layers stacking. Considering one-layer GCN, the 1st layer $\mathrm{H}^{(1)} \in \mathbb{R}^{N\times K}$ with K-dimensional node feature matrix is computed as

$$H^{(1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\,\widetilde{D}^{-\frac{1}{2}}\,\mathrm{X}\,W^{(0)}) \qquad (1)$$

Where $\widetilde{A}$ is adjacency matrix added with self-loop, $\widetilde{D}$ is the degree matrix of $\widetilde{A}$, let $\hat{A} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\,\widetilde{D}^{-\frac{1}{2}}$ be the normalized symmetric adjacency matrix, $W^{(0)} \in \mathbb{R}^{N\times K}$is a weight matrix. $\sigma(\cdot)$ is activation function, e.g. $ReLU(\cdot) = max(\cdot,0)$. As mentioned above, we can integrate larger scale of neighborhoods information by stacking multiple GCN layers like this:

$$H^{(l+1)} = \sigma(\hat{A}\,H^{(l)}\,W^{(l)}) \qquad (2)$$

Where $H^{(0)} = X$ and l denotes the layer number.

3.2 VCs Graph Convolutional Neural Network (vcGCNN)

We build a large graph which contains nodes of VCs so that co-investment among VCs can be simply modeled and graph convolutional neural networks can be easily adapted. The model based on graph convolutional neural networks for leading VCs detection, so called vcGCNN consist of three steps as describe:

Step 1: construct co-investment network.

Firstly, we have to define the co-investment relationship of VCs. We defined indicator I_{ij} as co-investment between VC *i* and VC *j*, where $t = 1,2,...,T$ is the investing timestamp. Accordingly, $I_{ij}(t) = 1$ indicates VC i and VC j have invested in the same company at the same time; $I_{ij}(t) = 0$ otherwise. After definition, an undirected co-investment network $G = (V, \mathcal{E})$ is constructed, where V is nodes set of VCs and $\mathcal{E}$ is edges set of relationships connecting VCs. For each edge (v_i, v_j), denote by $e_{ij} = \sum_{t=1}^{T} I_{ij}(t)$ the weights on the edge. The summary of the co-investment network is presented in Table 1.

Step 2: extract feature matrix.

In this work, we derive nine indicators related to investing scale and experience of VCs [21] from investment events dataset, to make up for the limitation of graph structure information. The status of VCs can be described by how many companies a venture capital firm has invested in (NoCom), which indicates whether the venture capital firm has sufficient assets and resources to invest or not. Similarly, how many times of investments overall (NoAll), how many industries has invested in (NoInd), how many periods has invested in (NoPer), how many countries has invested in (NoCou), how many provinces has invested in (NoPro), how many stages: initial stage (NoIni), expansion stage (NoExp) and seed stage (NoSee) has invested in all measure the scale and experience of VCs. The summary statistics of the nine feature indicators are presented in Table 2.

Step 3: build vcGCN model.

After building the co-investment graph and extracting feature matrix, we feed the graph and feature matrix into a two-layer GCN, where SoftMax classifier is used for classification:

$$f(X,A) = softmax(\hat{A}\ ReLU(\hat{A}\,X\,W^{(0)})\,W^{(1)}) \qquad (3)$$

Where *softmax* (x_i) is $\frac{1}{S}$*exp* (x_i) with $\mathrm{S} = \sum_i exp(x_i)$. We use stochastic gradient descent (SGD) to train weight parameters $W^{(0)}, W^{(1)}$, and cross-entropy error as loss function:

$$\mathcal{L} = -\sum_{d\in \mathrm{y}_D}\sum_{f=1}^{F} Y_{df}\ lnZ_{df} \qquad (4)$$

Where y_D is indices set of labeled VCs and F is the number of classes. In equation 3, $E_1 = \hat{A}\,X\,W^{(0)}$ implements the first layer convolutional operation, $E_2 = \hat{A}\ ReLU(\hat{A}\,X\,W^{(0)})\,W^{(1)}$ explore to the second layer of co-investment network. The overall vcGCN model is schematically illustrated in Figure 1.

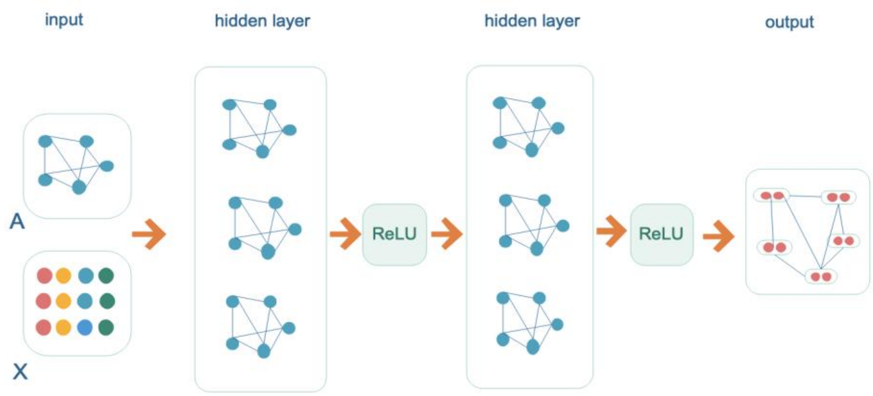

Figure 1. Graph convolutional neural network for leading venture capital firms identification

The vcGCNN model naturally integrate the connectivity patterns between VCs and feature attributes of VCs. We found that two-layer performs better than one-layer, but when we added more than two layers, the performances did not improve significantly.

4. EXPERIMENT

4.1 Baselines

- **Degree centrality (DC)**. Degree centrality is the sum of one node connected to other nodes [22,23]. A high DC score represents a node has a large number of neighbors.
- **Betweenness centrality (BC)**. Betweenness centrality refers to the number of shortest paths that a node appears between other nodes. In other words, this node is intermediary, and the nodes usually pass through it to reach other nodes.
- **Closeness centrality (CC)**. there must be shortest path from one node to the other, and the average length of these shortest paths is closeness centrality[22,23].
- **K-core centrality (K-core)**. K-core is the degree of a maximal subgraph in which all nodes have a degree of at least k [16] Also, it is formed by repeatedly deleting all nodes of degree less than k.
- **Eigenvector centrality (EC)**. The main idea of eigenvector centrality is that the influence of a node relatively depends on the influence of its neighbors. Computing eigen decomposition of adjacency matrix, and the eigenvector centrality of nodes is the corresponding eigenvector of the largest eigenvalue [15].
- **Chebyshev spectral CNN(ChebNet)**. Spectral-based methods have a solid mathematic foundation in graph signal processing, the key difference of which lie in the choice of filter [18]. Formally, define spectral convolutions on graph as the product of signal $x \in \mathbb{R}^N$ and filter g_θ: $g_\theta * x = U g_\theta U^T x$, $\theta \in \mathbb{R}^N$. ChebNet uses Chebyshev polynomials of diagonal matrix of eigenvalues as filter [19]. i.e. $g_\theta(\wedge) \approx \sum_{k=0}^{K} \theta_k T_k(\widetilde{\wedge})$. Where U is the matrix of eigenvectors of the normalized graph Laplacian $L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U \wedge U^T$, and $\widetilde{\wedge} = \frac{2}{\lambda_{max}} \wedge - I_N$. The Chebyshev polynomials are defined as $T_k(x) = 2x T_{k-1}(x) - T_{k-2}(x)$ which is recursively, with $T_0(x) = 1$ and $T_1(x) = x$. Going back to the definition of spectral convolutions on graph, we now have:

$$g_\theta * x \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L})\, x \qquad \tilde{L} = \frac{2}{\lambda_{max}} L - I_N \tag{5}$$

- **First order of ChebNet (1st order model)**. Set *K*=1, equation 5 is linear. In the linear formulation we further approximate $\lambda_{max} \approx 2$. Under these approximations, equation 5 simplifies to:

$$g_\theta * x \approx x\theta_0 + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} x \theta_1 \tag{6}$$

Then, we set a single parameter $\theta = \theta_0 = \theta_1$, this leaves us with a simpler expression:

$$g_\theta * x \approx \theta(I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}})x \tag{7}$$

4.2 Dataset

In china, major venture capital databases (ChinaVenture, Zero2IPO, and Venture Capital Research Institute's annual reports) release public investment and related index data, and we have collected all venture capital investment events from 1993 to 2014. Each investment event indicates that a VC has invested in a company. At the same time, each investment event records when and where the venture capital company invested in which company, and lists the industry to which the company belongs and the investment period (initial stage, expansion stage, seed stage) at that time.

We construct a large co-investment network from collected dataset, which contains 8,680 nodes and 14,789 edges. The maximum value of the edge weight is 272, which indicates that two VCs have the highest co-investment of 272 times between 1993 and 2014. The summary of the co-investment network is presented in Table 1, and the summary statistics of the nine feature indicators are presented in Table 2.

Table 1. The summary of the co-investment network

#Nodes	#Edges	#Classes	Weights on Edges			
			Max	Min	Mean	Std
8680	14789	2	272	1	9.0587	14.8933

Table 2. The summary statistic of the nine feature indicators

Indicators	Mean	Std	Median	Min	Max	Skew	Kurtosis
NoC	2.4914	12.0964	0.0000	0.0000	463.0000	17.6994	498.4426
TNI	3.1347	16.6226	0.0000	0.0000	628.0000	18.1528	499.9633
NoI	1.5173	4.5399	0.0000	0.0000	101.0000	7.2042	78.9170
NoP	0.7089	1.1830	0.0000	0.0000	5.0000	1.8159	2.6722
NoCoun	0.4534	0.7717	0.0000	0.0000	10.0000	3.2962	20.8392
NoPR	0.9250	2.2605	0.0000	0.0000	28.0000	5.0498	34.6475
NoSI	0.8275	5.3147	0.0000	0.0000	215.0000	21.3126	686.2657
NoSE	1.1848	6.4657	0.0000	0.0000	298.0000	21.1435	709.4396
NoSS	0.3808	3.1130	0.0000	0.0000	138.0000	22.8652	733.3084

4.3 Settings

For vcGCNN, we set the learning rate as 0.01, dropout rate as 0.5, L_2 loss weight as 5e-4, hidden units as 16, validation set as 10% of training set selected randomly. Following the settings of GCN, we trained vcGCNN with Adam [24] for a maximum of 200 epochs. For ChebNet based methods, we used the same parameter settings but the different filter.

4.4 Accuracy analysis

We have carried out Delphi inquiry interviews in April 2013, and obtained reference label of leading VCs. Before the interview, we calculated K-core centrality of all the VCs and listed 908 VCs based on the K-core score. Then we interviewed four experts who are familiar with the Chinese venture capital market. They are asked to choose the most influential VCs from the list, and tick out "yes" or "no" for all the listed VCs. Among the four experts, one is the leader of a VC research institute in the Chinese central government, and the other three are CEOs of large foreign or domestic VCs. Finally, we selected 42 of the most influential VCs from the checklist as leading VCs, and the others are taken to be not leading VCs.

4.5 Test Performance

For centrality measures methods, we separately consider top-30, top-40, top-50 ranking as leading VCs, the test accuracy of each centrality measures method is presented in Figure 2. As figure illustrate, the accuracy of centrality measures is near 68%, which is not good enough for leading VCs identification.

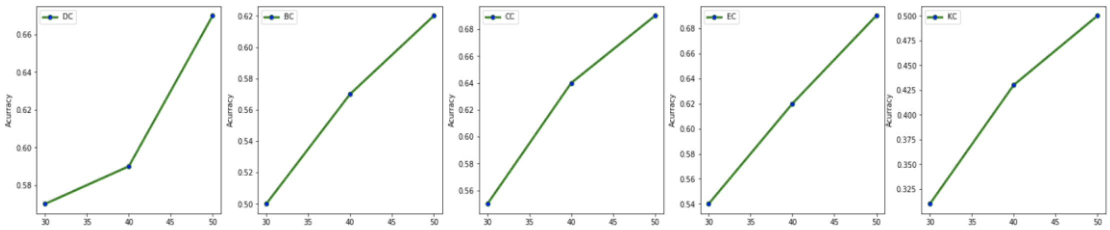

Figure 2. Test accuracy of each centrality measures method

We compare ChebNet with K is 2 or 3, and different per-layer propagation model on the co-investment graph. Following the settings described in the section4.3, the results are summarized in Table 3. The test accuracy result of our vcGCNN model is in bold. Reported numbers denote mean test accuracy for 10 repeated runs. In case of multiple variables Θ_i per layer, we impose L_2 regularization on all weight matrices of the first layer.

Table 3. Comparison of propagation models

Description		Propagation model	Accuracy
ChebNet (Eq. 5)	K= 2	$X\Theta_0 + \tilde{L}X\Theta_1 + (2\tilde{L}^2 - 1)X\Theta_2$	0.9325
	K= 3	$X\Theta_0 + \tilde{L}X\Theta_1 + (2\tilde{L}^2 - 1)X\Theta_2 + (4\tilde{L}^3 - 3\tilde{L})X\Theta_3$	0.9171
1^{st}-order model (Eq. 6)		$X\Theta_0 + D^{-\frac{1}{2}}AD^{-\frac{1}{2}}X\Theta_1$	0.9238
single parameter (Eq. 7)		$(I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}})X\Theta$	0.9289
vcGCNN (Eq. 2)		$\boldsymbol{\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta}$	**0.9346**

Table 3 presents test accuracy of each spectral-based model. vcGCNN outperforms all baseline models on co-investment network, which demonstrates the effectiveness of the proposed approach on investment events datasets. For more performance analysis, we note that other propagation models also perform well on leading VCs identification. This is likely due to the fact that spectral graph convolutional network is very suitable and efficient for nodes classification on graph-structure data, especially on co-investment graph of VCs. The performance of different spectral-based model is delicate, after all, our proposed model is better.

4.6 Label Rate Sensitivity

In order to evaluate the effect of the size of the labeled data, we tested several best performing models with different proportions of the training data. Classification accuracies on testing data with label rate as 1%, 5%, 10%, 20% and 40% are reported in Figure 3. Result shows that vcGCNN can achieve higher test accuracy with labeled nodes added and performs well with limited labels. These encouraging results suggest that vcGCNN can propagate VCs label to the entire graph well and the graph preserves global co-investment information. Mention again，1^{st} ChebNet and single parameter model which are similar to vcGCN also performs well with different proportions of training data.

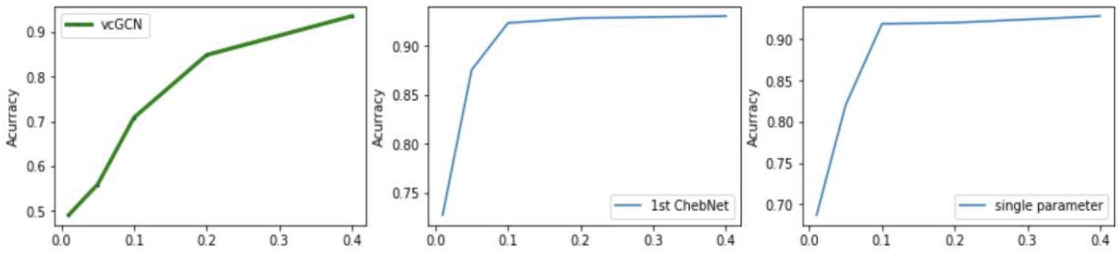

Figure 3. Test accuracy by varying training data proportions

4.7 Co-investment Network Visualization

We give an illustrative visualization of the co-investment network. Figure 4.1 shows the visualization of 8680 venture capital firms with 14789 co-investment relationship. As we can see, there are some of isolated nodes around the graph, we then remove less than 5 times co-investment relationship in the network, which leaves 1123 nodes and 6607 edges, the new graph is showed in Figure 4.2.

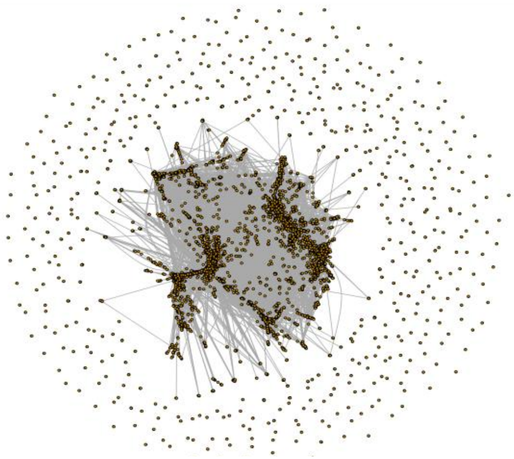

Figure4.1. The co-investment graph

Figure4.2. Subgraph where weights larger than 4

5. CONCLUSION

To sum up, we propose a new method for leading VCs identification termed VCs Graph Convolutional Neural Networks (vcGCNN). A large graph is built for investment events dataset, so that the problem of leading VCs identification is turned into the problem of node classification. vcGCNN has good performance when capturing graph structure information and learning from limited labels. A simple two-layer vcGCNN demonstrates promising results by outperforming numerous centrality measures methods and typical spectral GNN methods on leading VCs identification. Additionally, there are some interesting directions for future work, including improving the classification performance using different propagation rules, tending to link prediction to search for the most feasibility co-investment partner, and developing unsupervised model for unlabeled graph-structure data.

REFERENCES

[1] 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 firms. Asia Pacific Journal of Management.

[2] Barabási A L. (2005). Taming complexity. Nature physics.

[3] Wilson R. (1968). The theory of syndicates. Econometrical: journal of the Econometric Society, 119-132.

[4] Tykvová T. (2007). Who chooses whom? Syndication, skills and reputation. Review of Financial Economics, 16(1): 5-28.

[5] Hochberg Y V, Ljungqvist A, Lu Y. (2007). Whom you know matters: Venture capital networks and investment performance. The Journal of Finance, 62(1): 251-301.

[6] DUFOUR D, NASICA E, TORRE D. (2011). Optimal syndication choices in venture capital investment: understanding the role of skills and funds providers.

[7] 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.

[8] Aral S, Walker D. (2012). Identifying influential and susceptible members of social networks. Science.

[9] Kipf, T. N., and Welling, M. 2017. (2017). Semi-supervised classification with graph convolutional networks. ICLR.

[10] Hou B, Yao Y, Liao D. (2012). Identifying all-around nodes for spreading dynamics in complex networks. Physica A: Statistical Mechanics and its Applications.

[11] Szolnoki A, Xie N G, Ye Y, et al. (2013). Evolution of emotions on networks leads to the evolution of cooperation in social dilemmas. Physical Review E.

[12] Zhong L, Gao C, Zhang Z, et al. (2014). Identifying influential nodes in complex networks: A multiple attributes fusion method. International Conference on Active Media Technology. Springer.

[13] Sabidussi G. (1966). The centrality index of a graph. Psychometrika, 31(4): 581-603.

[14] Freeman L C. (1978). Centrality in social networks conceptual clarification. Social networks, 1(3): 215-239.

[15] Bonacich P. (2007). Some unique properties of eigenvector centrality. Social networks, 29(4): 555-564.

[16] Kitsak M, Gallos L K, Havlin S, et al. (2010). Identification of influential spreaders in complex networks. Nature physics, 6(11): 888.

[17] Page L. (1998). The pagerank citation ranking: Bringing order to the Web. Technical report. Stanford Digital Library Technologies Project.

[18] Wu Z, Pan S, Chen F, et al. (2019). A comprehensive survey on graph neural networks. arXiv preprint arXiv:1901.00596.

[19] Defferrard M, Bresson X, Vandergheynst P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems. NIPS.

[20] Levie R, Monti F, Bresson X, et al. (2018). Cayleynets: Graph convolutional neural networks with complex rational spectral filters. IEEE Transactions on Signal Processing, 67(1): 97-109.

[21] Yang H, Luo J D, Fan Y, et al. (2019). Using weighted k-means to identify Chinese leading venture capital firms incorporating with centrality measures. Information Processing & Management, 102083.

[22] Gao S, Ma J, Chen Z, et al. (2014). Ranking the spreading ability of nodes in complex networks based on local structure. Physica A: Statistical Mechanics and its Applications, 403: 130-147.

[23] Huang C Y, Fu Y H, Sun C T. (2015). Identify influential social network spreaders, IEEE International Conference on Data Mining Workshop. IEEE, 562-568.

[24] Kingma D P, Ba J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.