Analysis of Chinese Venture Capital Networks

Liz Guo, Weini Yu, Junlin Liu

December 9, 2018

1 Introduction

Venture capital (VC) firms are those who provide funds and other resources to startups in order to help them grow from scratch to successful companies [1]. Venture capital investments can be profitable, especially when companies they invest in finally go public. On the other hand, they are highly risky because the overall survival rate for startups is extremely low [2][7]. To reduce risk, venture capitals usually include a few startups in their portfolio, and they tend to invest together with other VC firms rather than individually. All of these activities form different networks of venture capital investments that we can study [3][8].

While venture capital has been an industry for decades in the United States, it is still relatively new in China. There were few venture capital firms only ten years ago. Now, it has become an industry with hundreds and thousands of venture capital firms and the total asset under management (AUM) has reached trillions of RMB. For such a young industry, still minimal research has been conducted from the perspective of network analysis.

Therefore, in this project, we would like to analyze the networks of Chinese venture capital firms and their investments. We conduct two types of analysis, static and dynamic, on both undirected and directed networks to obtain a comprehensive understanding. In static network analysis, we focus on community detection, as well as motif analysis and node impact evaluation. In dynamic network analysis, we study the evolution of Chinese venture capital networks.

2 Related Works

In the past, some interesting work has been conducted on venture capital networks. Jin et al. [12] study the characteristics of venture capital network in China, compare their operations with those of the western VC firms, and focus on their relationship with regional economy including industrial structure, and employment. [14] examines the relationship and organizational structure established from VC syndication and finds that better networked VCs have significantly better fund performance measured by the proportion of investments that are successfull IPO exits or sales to another company. Xue et al. [1] study the evolution of Chinese VC investment networks and how that affects performance of those VC firms through a linear regression model. They conclude that movements between different communities have positive impact on performance of venture capital firms in terms of the number of IPO exits and internal rate of return (IRR).

To understand venture capital networks, one of the most relevant features we need to learn is community structure. Traditional community detection algorithms such as spectral methods [4], Louvain algorithm or node (graph) embeddings can be applied and we explain in detail in Section 4.1.2. There are other interesting approaches based on deep learning [5] or cylic patterns [6] to find clusters in graph.

Predicting investment behavior is another popular topic. Liang et al. [3] studies the funding investors investment in companies based on social relationships. We won't explore it in this project but it could be future work.

3 Data and Representations

3.1 Data Collection and Statistics

Our data is obtained from crunchbase.com, a well-known commercial database of venture capital investments. For the purpose of our project, we filter out venture capital firms that are not headquartered in China. Out of these 1555 VC firms, we discard nearly 1000 trivial VC firms who have only made one investment in total (or only one record has been collected), and are left with 512 VC firms and 9344 transactions (investments) associated with them.

To get an overview of what the data looks like, we perform some basic statistical analysis and present the summary below. Figure 11 in appendix shows the distribution of the number of investments venture capital firms have made. We see that most venture capital firms have made less than 30 investments or so and very few firms have made hundreds of investments in total. Figure 12 shows most startups have received no more than 2 investments from venture capital firms, and the number of startups receiving more than 5 investments is really low.

We then take a look at the distribution of funding rounds, shown in Table 1. 80% of all funding rounds are between seed stage and series G. For simplicity and interpretability, we will only adopt these 7000+ transactions to build our directed network later. The distribution of years when these funding rounds happen has been inspected as well, shown in Table 2. Since the number of investments before 2010 is very low, we are going to lump them together when studying the evolution of these networks later.

Funding Round	Number	Ratio
Seed	1011	10.82 %
Angel	503	5.38 %
Series A	2613	27.96 %
Series B	1991	21.31 %
Series C	1005	10.76 %
Series D	415	4.44 %
Series E	130	1.39 %
Series F	44	0.47 %
Series G	21	0.21 %
Others or Unknown	2026	21.70 %

Table 1: The number and ratio of funding rounds

Year	# Funding Rounds
2018	2276
2017	1607
2016	1284
2015	1160
2014	926
2013	486
2012	283
2011	339
2010	215
2009 and before	768

Table 2: The distribution of years in which funding rounds happen

3.2 Network Representations

For our project, more than one type of network can be constructed for a comprehensive analysis. For example, we can run community detection algorithms on both unweighted and weighted networks and compare the results. We can also extend an undirected network to a directed one where we can discover significant motifs and study patterns. Furthermore, we can analyze not only networks of investors, but also those of startups. Thus, we define notations for the unweighted networks as follows:

G_{iu}, the undirected network of investors. Each node represents an investor. An edge exists between two nodes if two investors have invested in the same startup.

G_{id}, the directed network of investors. If investor A invests in a startup in one round and then investor B invests in the same startup in the next round (E.g. investor A invests in startup X in Series D and investor B invests in startup X in Series E), then a directed edge exists pointing from node B to node A.

G_{su}, the undirected network of startups. Each node represents a startup. An edge exists between two nodes if two startups share a common investor.

4 Static Network Analysis

In static network analysis, we consider all investment activities in history as a whole. For network G_{iu}, we provide comprehensive measurements and description and present highly interpretable results of community detection using different algorithms. For G_{id}, we find statistically significant motifs that reveal interesting investment patterns, and calculate nodes' PageRank scores which can be an alternative way of evaluating the venture capital firms. We also give a brief summary of the community detection results on G_{su}.

4.1 Analysis of Network G_{iu}

4.1.1 Network Measurement

There are 512 nodes and 3232 edges in G_{iu}. It is composed of a very large weakly connected component (WCC) with 439 nodes and 73 isolated points. We calculate the distributions of node degrees and clustering coefficients. Figure 1 shows that the distribution of node degrees agrees with the power law. In figure 2, we see clustering coefficients of nodes are high in general. The average clustering coefficients of G_{iu} is 0.42. A heuristic algorithm is adopted to approximate the size of the largest clique in G_{iu} and the result is 13, which is a little lower than our expectation. We also check the distribution of shortest paths between all node pairs and find that lengths of most shortest paths are less than 5. The diameter of the whole network is only 6, which is consistent with the findings in the famous Small-world Experiment. We then generate hundreds of configuration models from the graph and the average diameter of them is 5.4, showing that the diameter of our network is intrinsic in the distribution of its node degrees.

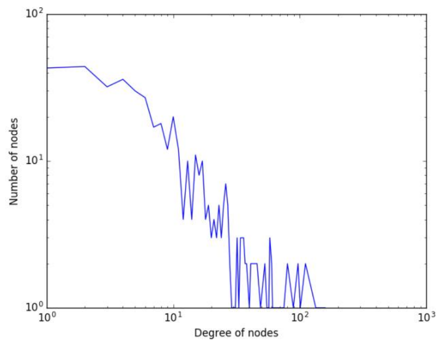

Figure 1: Distribution of node degrees

The facts above together show the compactness of Chinese venture capital networks built from their co-investment activities. The young and fast-developing Chinese venture capital industry is indeed a "small world".

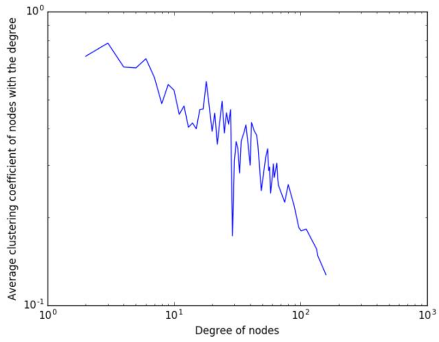

Figure 2: Distribution of node clustering coefficient

4.1.2 Community Detection

Some venture capital firms may have similar investment styles or themes. Those who do may conceivably form a community in the network representation. In this section, we try to find those communities and their characteristics. We use different community detection methods on G_{iu} including Louvain algorithm, spectral clustering and node2vec clustering. We also run Louvain algorithm on weighted G_{iu}. Since the clustering task on isolated nodes is trivial, we discard all of them and only detect communities on the WCC.

4.1.2.1 Louvain Algorithm

The Louvain algorithm greedily maximizes modularity Q which is defined as

$$Q(G,S) = \frac{1}{2m}\sum_{s\in S}\sum_{i\in s}\sum_{j\in s}(A_{ij}\frac{k_i k_j}{2m})$$

where S are the partitions, m is the number of edges of graph G, $A_{ij}=1$ if an edge exists between nodes i and j otherwise 0, k_i and k_j are the sum of edges attached to node i and j respectively.

The algorithm starts with each node in its own distinct community. In the partition phase, it iteratively tries to move each node i to the community of some neighbor j that yields the largest modularity gain, until no movement can be made. In the restructuring phase, it contracts the partitions from the partition phase into super-nodes and updates the edges accordingly. The two phases run in turn until the community configuration does not change anymore.

The Louvain method clusters the 439 nodes in the WCC into 9 communities, with a modularity of 0.2416. For comparison, the configuration model with the same degree sequence is partitioned into 10 communities with a modularity of 0.1662 which is much lower. The degree distribution of nodes in each community is shown in Figure 3. We see each community has a few supernodes with very high degrees followed by more smaller nodes. Empirically, this means each community is led by a few bigger VCs with larger number of investments.

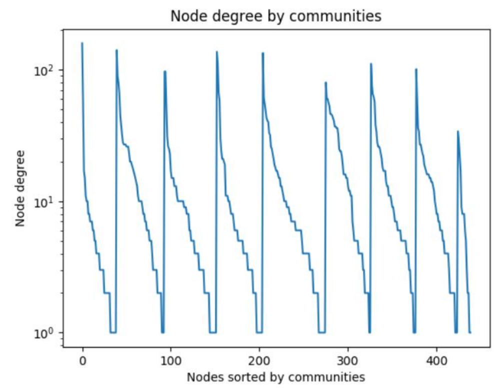

Figure 3: Distribution of node degrees sorted by communities

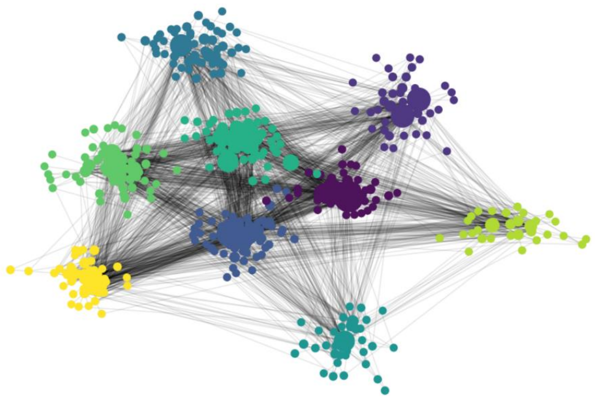

Figure 4: Communities detected by Louvain algorithm. The node size is proportional to the number of investments the VC has made.

Result Interpretation To further investigate characteristics of these communities, we inspect representative venture capital firms of each community, shown in Table 3. The results obtained by Louvain algorithm are fairly good in terms of interpretability. For more than half of the communities, it is easy to see characteristics shared by the venture capital firms in them.

For example, community 2 has Tencent, Alibaba, Baidu, Ant Financial (a subsidiary of Alibaba), which are exactly the Big Three in the Chinese Internet industry. That means, the 3 CVCs (Corporate Venture Capital) share similar investment style and focus on similar tracks and projects.

Community 1 consists of many biotech companies and venture capital firms focusing on biomedicine industry. Lilly Asia Ventures is the venture capital department of Lilly, a large company in pharmaceutical industry. Sequoia and Qiming are the two Chinese venture capital firms that heavily invest in biomedicine industry.

Venture capital firms in community 4 focus on the earliest stage investments including seed and angel stages. Those in community 6 focus on early stages too, but they are not only capital providers but also incubators for startups. In fact, Sinovation Ventures, founded by Kaifu Li, a famous Chinese entrepreneur, is the first startup incubator in China.

The common characteristics of venture capital firms in community 0 is that they are major players in late stage investments. Community 5 has companies or venture capital firms that are closely related to Jun Lei, the founder of Xiaomi, a famous Chinese smartphone manufacturer.

4.1.2.2 Spectral Clustering

We also explore k-way spectral clustering algorithm [9] which we implemented from scratch to detect communities on G_{iu}. In k-way spectral clustering method, each node is represented by a k-dimensional vector derived from k eigenvectors of the Laplacian matrix of G_{iu}. Then we cluster the nodes by their k-dimensional vector representation.

The number of clusters k is the most critical parameter which is usually set manually. Zelnik-Manor et al [10] discussed two approaches to find k, which we adopt in our implementation. The first and more intuitive approach is to analyze the eigenvalues and look for the k value that maximizes the eigengap $\Delta k = |\lambda_k - \lambda_{k-1}|$. We plot the first 16 eigenvalues of our graph Laplacian matrix

Community	Size	Representitive VCs
0	50	IDG Capital, Shenzhen Capital Group, Legend Capital
1	40	Sequoia Capital China, Qiming Venture Partners, Lilly Asia Ventures
2	70	Tencent Holdings, Alibaba Group, Baidu, Ant Financial
3	52	Matrix Partners China, Node Capital, FREES FUND
4	34	ZhenFund, PreAngel, Decent Capital
5	59	Shunwei Capital, Morningside Group, Xiaomi
6	55	Sinovation Ventures, Microsoft Accelerator Beijing, Cherubic Ventures
7	30	Source Code Capital, Bertelsmann Asia Investment Fund
8	49	SB China Venture Capial, Fortune Venture Capital

Table 3: Communities and representative venture capital firms detected by Louvain algorithm.

which is shown in Figure 5. It suggests the biggest magnitude drop of the eigenvalues is when $k = 4$.

The second approach provides a more theoretical justification which relies more on eigenvectors. The cost function is defined as

$$J = \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{Z_{ij}^2}{M_i^2}$$

where C is number of groups, Z is matrix obtained after rotating the eigenvector matrix, and $M_i = max_j Z_{ij}$. Due to page limit we can't describe in detail but by minimizing this cost we get the best k which is also shown in Figure 5.

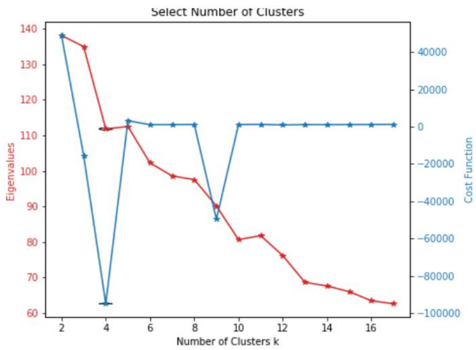

Figure 5: Select number of clusters using eigenvalue gap and cost function. When k=4 we find the biggest eigenvalue gap as well as the minimum cost.

These two different selection methods yield the same number of clusters. But the result delivered is quite different from that of Louvain algorithm and is less interpretable. In addition, the distribution of community sizes of k-way spectral clustering is far more skewed than Louvain algorithm. This may be because k-way spectral clustering does not explicitly consider the balance of community size as an optimization target.

4.1.2.3 Node2vec

The node2vec algorithm can learn node feature representation in networks while preserving local neighborhoods of nodes. To sample appropriate neighborhoods, this algorithm proposes two search strategies, breadth-first sampling (BFS) and depth-first sampling (DFS). The neighborhoods sampled by BFS lead to embeddings that correspond to structural equivalence, while those sampled by DFS reflect communities based on homophily.

In order to discover communities, we use DFS and set return parameter p=1, in-out parameter q=0.5, then run node2vec to learn feature representation in a 128-dimensional feature space. It learns macroscopic view of the network neighborhoods. Then we apply k-means method to cluster nodes into 6 communities.

We also check several representative venture capital firms in each community. Compared to results given by spectral clustering, they are more interpretable as node2vec does preserve some distinctive communities, though still not as good as Louvain algorithm. As shown in Table 4, we get one community of venture capital firms that focuses on biomedicine industry and one that focuses on blockchain applications and cryptocurrencies.

Community 3	Lilly Asia Ventures, Decheng Capital 3E Bioventures, BioVeda China Fund
Community 4	Node Capital, BlockVC Fenbushi Capital, Bitmain

Table 4: Two distinctive communities with representative VC firms given by node2vec algorithm.

Since the original node embeddings have 128 dimensions, we apply PCA to reduce the number of dimensions while keeping most information. Figure 6 is a scatter-plot of the first and second principal component. Different colors correspond to different communities. The first two principal components account for 74% of the variation so it's able to set communities apart nicely.

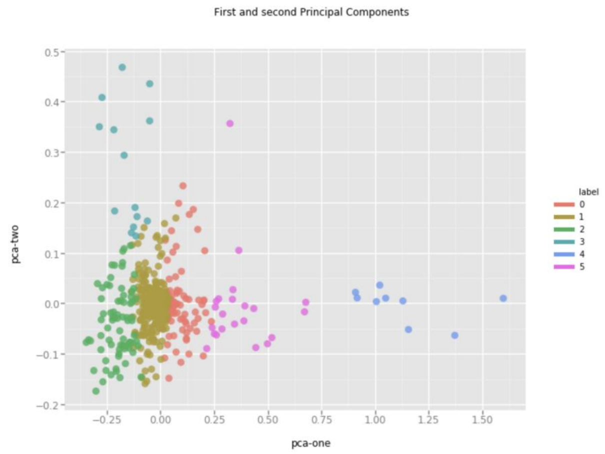

Figure 6: Community visualization based on node embeddings.

In Figure 6, we see the distribution of community sizes is also skewed. One large community (community 1) consists of nearly half the largest venture capital firms in the market. Community 3 and 4 are the most distinctive in this figure. Other communities are closer to each other and thus difficult to set apart completely. We think this is why node2vec algorithm is not as good as Louvain on our task.

4.1.2.4 Louvain on Weighted G_{iu}

To see how the result of community detection change when assigning weights to edges, we also investigate the community structure on weighted G_{iu} using Louvain algorithm. The edge weights are modeled in two different ways: common neighbors and Jaccard Index.

Common Neighbors G_{iu_cn} In this graph, the edge weights between two investors are the number of common startups the they have invested in.

$$w(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

where $\Gamma(x)$ is the set of startups that investor x invest in.

On this weighted network, we again get 9 communities with a slightly lower modularity of 0.2126 compared to the unweighted graph indicating worse community structure on weighted network.

Jaccard Index G_{iu_jc} Because modeling edge weights using common neighbors tends to overweight larger investors with more investments, we want to mitigate this effect by using Jaccard Index instead. The edge weight between two nodes are defined as the number of common startups the two investors invest in over the union of all the startups they invest in.

$$w(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

On this weighted network, Louvain gives 21 communities with a higher modularity of 0.5513 than unweighted one. Though the modularity is higher, the results have a similar problem to that of node2vec algorithm. That is, there is a very large community containing 90% of the largest venture capital firms in terms of the number of investments they make, and only community of bio-ventures and crypto-ventures can be clearly detected.

The number of communities we get is much more than unweighted one as there are many small-sized communities (2 or 3) where the Jaccard weights between the nodes in these small communities are large. They are investors who make just a few investments together. The large weights between them prevent them from being merged into other larger communities due to the way Louvain maximizes modularity.

To dive deeper into why it gives a much higher modularity, we iteratively merge the smallest communities and see how the modularity changes. As shown in Figure 7, the modularity only drops below the level of the unweighted network after 18 iterations (when only 2 communities left). This is surprising because this 2 communities clustering should be much worse than the 9 communities we get before.

So we experiment re-calculating modularity of the unweighted network using assignment given by G_{iu_jc}, and modularity of G_{iu_jc} using assignment given by the unweighted network. That yields modularity of 0.1875 and 0.4402 respectively. This tells us that the

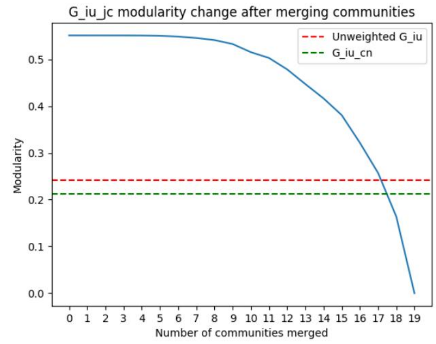

Figure 7: G_{iu_jc}'s modularity change when iteratively merging the two smallest communities into a single one

increase in modularity score on this weighted network is not a result of actually better clustering but is simply dominated by the Jaccard edge weights.

4.2 Analysis of Network G_{id}

Now we approach the investor network from a different perspective and study G_{id}. There are a few ways to add direction to edges. One is that if there are lead investors in a funding round, we can add directed edges from other investors to the lead investors in the same funding round. However, the dataset does not have much information about lead investors and in many funding rounds there are no lead investors at all, which will make the network more sparse. Therefore, we adopt the method described in section 3.2. This way, an edge's direction indicates the successfulness of an investment a venture capital firm has made in the last funding round of a startup to some extent.

4.2.1 Motif Analysis

In this section, we are going to find the significance of different motifs in G_{id}. We focus on motifs formed by three nodes and directed edges connecting them. Figure 10 in appendix shows all possible motifs.

Method Here we adopt a commonly used method to conduct our analysis. First, we run ESU algorithm to enumerate all subgraphs formed by 3 nodes, check which motif it is and then increment the counter of the motif. Then we rewire the edges to get several random networks. We calculate the average number of times each motif appears in these random networks and compare them with that in the original network (G_{id}) using z scores which is given by the following formula

$$Z_i = \frac{N_i^{real} - \bar{N}_i^{rand}}{std(N_i^{rand})}$$

where N_i^{real} is the number of times the i-th motif appears in our network. $\bar{N}_i^{rand}$ and $std(N_i^{rand})$ are average and standard deviation of the number of times it appears in rewired random networks.

Result The z score vector we get is [-6.14, -6.68, 0.66, -7.40, -3.64, -1.45, 3.93, **7.09**, -4.29, -5.97, 1.65, 2.30, **6.89**]. Typically, motifs with the largest positive values are considered as significant. So motif 8 and 13 are statistically significant motifs in G_{id}. These two motifs share a characteristic: the edges between two nodes are bidirectional. The implication of this pattern is that, in actual investment, an investor A may invest in a startup in a round that follows the round an investor B participates, but it may invest in another startup a round ahead of investor B.

4.2.2 Evaluation of Node Impact

In this section, we adopt PageRank to evaluate nodes in our network. The motivation here is that we want to give an alternative way to assess the impact of venture capital firms. There are two common approaches. One is performance-based, i.e. how many unicorns it has invested in? Or more directly, what the internal rate of return (IRR) of the fund? Another is scale-based, i.e. how much is the asset under management (AUM) of the venture capital firm? Or how many investments it has made? Some of these metrics are good but not sufficient, while some are usually confidential. Therefore, our method looks at how good a venture capital firm is or how successful their investments are from another perspective, based on the way we build G_{id} as discussed before.

PageRank Each node i in the network has a score $R(i)$, which can be iteratively computed by

$$R(i) = d \sum_{j \in N(i)} R(j) \frac{G_{id}(j,i)}{\sum_k G_{id}(j,k)} + \frac{1-d}{|V(G_{id})|}$$

where d is the teleport factor, $N(i)$ is set of neighbors of i that has an edge pointing to i, $G_{id}(j, i)$ is the weight of the edge from j to i, and 0 if there is not such an edge.

The resulting score can be seen as the impact score of a venture capital firm. We output the top 20 venture capital firms (call it List A) and compare it to the list ranked by the number of investments a venture capital firm makes (call it List B). We present just a few examples of our findings below.

Finding 1 The top 5 investors in List A are exactly the top 5 in List B. This is not surprising. But the 6-th investor in List B, ZhenFund, only ranks 18 in List A. ZhenFund is well-known in China for its investments in young entrepreneurs especially fresh graduates. Its founder, Xiaoping Xu, is very active in media and social events, which makes ZhenFund very popular among young people. However, according to our PageRank analysis, its investments are not as good as its popularity in young people.

Finding 2 Another interesting example is Ant Financial, which only ranks 45 in List B but ranks 9 in List A. That means it does not make too many investments, but the quality of its investments is rather good and initiative, which gives it a high impact score.

4.3 Analysis of Network G_{su}

We also perform community detection on the undirected startup network G_{su}, which is composed of 4914 nodes and 349041 edges. Using the Louvain algorithm, we get 91 communities with a modularity of 0.6199, compared to 0.0496 for the configuration model. Although 0.6199 modularity score does indicate good community structure, the clustering is less informative. The community size distribution is shown in in Figure 8. Only 18 out of 91 (20%) communities have more than 20 nodes. 67 communities have size no more 7 and each of these startup communities is associated with only one VC firm. In general, the clustering simply shows the groups of startups who have gotten funding from the same investors.

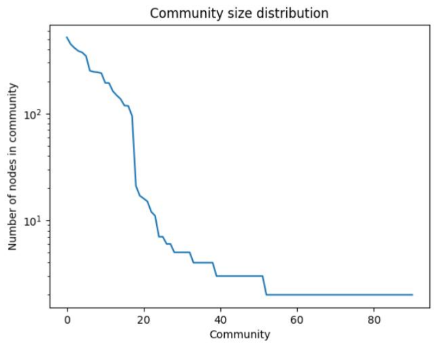

Figure 8: Community size distribution of G_{su}

5 Dynamic Network Analysis

Venture capital industry in China is constantly evolving. In this section, we study the changes of the networks over time on the undirected investor network G_{iu}. We concentrate on changes of communities since this is, in our opinion, the most interesting and informative task given the way we define our networks.

5.1 Network Division

There are many choices of how to build timestamped networks. First, we decide to divide the whole network by year. Thus we have an independent network for each year. A special case is that we combine networks before 2010 into a single one due to scarcity of transactions. We do not divide the network by two or three years because of its wide span, nor do we smooth these networks by linearly combining adjacent ones because we want to study what happen exactly each year. In general, the sizes of divided networks increase over time. Their properties are similar to the entire network except clustering coefficients are moderately lower.

5.2 Community Evolution Analysis

There are four classes of methods to track dynamic community evolution[11]. The method we adopt is doing independent community detection on each network for simplicity and interpretatbility.

Overall Trend After getting communities of each network, we take a look at their modularities, which is shown in Figure 9. The figure shows a descending trend but overall relatively high modularity (around or above 0.4), indicating good community structure in these networks. Meanwhile, an ascending trend of the number of communities can be noticed as well. Before 2015, the number of communities detected waves slightly around 10. After that, the number goes up to around 16. We think these two trends result from increasing network sizes.

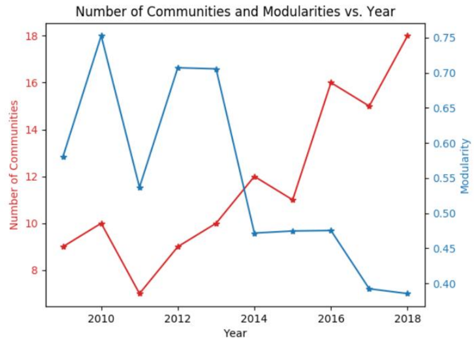

Figure 9: Modularity and the number of communities over time

Pattern Finding To further investigate characteristics of the communities, we again inspect them one by one. We output **contributing startups** for every community. A contributing startup for a community is one that receives funding from at least two venture capital firms in the community. This way, we can find out all startups that contributes to generation of edges in a community and then further figure out the pattern of the community.

Different from results in section 4.2, characteristics of communities on divided networks have less to do with investment stages or non-investment connections between representative VC firms. Instead, a typical community within a single year usually consists of 1 to 3 major VC firms and several other ones that co-invest with them. For instance, we have a community in 2011 shown in Table 5 and its contributing startups in Table 6.

1 Sequoia Capital China	8 Sinovation Ventures
4 Matrix Partners China	19 Gobi Partners
47 Green Pine Capital	57 Zero2IPO Ventures
70 Oriental Fortune	58 Capital Today
171 Taishan Invest AG	309 Zero2IPO Capital

Table 5: A community of the network in 2011. Numbers are IDs we give to venture capital firms.

Venture capital firms in Table 5 do not seem to share features. Nor do startups in

MoboTap (1, 4)	ihush.com (4, 58)
Camera360 (19, 4)	Umeng (4, 8)
Tuniu (1, 19)	Jiuxian (1, 70)
Kingnet (4, 57)	Hoolai Games (1, 309)
Ganji (1, 58)	Youbei Game (57, 47)
VIPstore.com (171, 47)	Doodle Mobile (8, 57)

Table 6: Contributing startups of the community. Numbers in brackets are IDs of VC firms that invest in the startup.

Table 6 do similar businesses (their businesses range from online games to retail to travel service). However, what can be noticed is that Sequoia Capital China (#1) and Matrix Partner China (#4) account for many of the edges in this community. Therefore, we can think of them as **dominant nodes** in the this community.

Interpretation of community changes We represent each community in brackets with IDs of its dominant venture capital firms or a letter indicating its industry in Table 7 and omit tiny communities. We can see that most of the communities are dominated by largest venture capital firms (those with ID less than 10). For example, the largest three, IDG Capital (#0), Sequoia Capital China(#1) and Qiming Venture Partners(#2) almost always dominate a community every year.

2009	(0,1)(2)(7)(9)(50)
2010	(0)(1)(9)(50)(57)
2011	(0)(1,4)(2,10)(15,45)
2012	(0,2)(1)(5)(8,15)
2013	(0,23)(1)(2,8)(4,11)(5,10)(38)
2014	(0,2)(1)(3)(4,8)(5)(10)(9)(16,19)
2015	(0)(1,22,33)(2,9,11)(3,13)(4,10,21)(5,16)
2016	(0,13)(1,3)(2,8,9)(4,5,10,23)(11,32)(21)(b)
2017	(0)(1,3,5,8)(2)(4)(7)(13,25)(41,61)(b)
2018	(0,1)(4,10)(3)(5,21,27)(8)(13)(20)(55)(c)

Table 7: Communities on networks of different years. Letter ‘b’ stands for bio-medicine and ‘c’ for cryptocurrencies.

In more recent years, the number of dominant investors in a community increases. This is reflective of a trend of the market - as both money and participants boom in the market, investors tend to compete for promising projects or unicorns which result in more co-investments, while in earlier years, investments are more exclusive.

Dominance can also be seen as an indicator of impact of a VC firm in the market. Before

2010, Shenzhen Capital Group (#7), a state-owned venture capital firm and one of the few large funding providers, has great impact since it is a dominant node in a community. After 2010 it no longer appears as a dominant node in communities until 2017. The same problem occurs to Fortune Venture Capital (#50), another state-owned venture capital firm in China. In contrast, we can see Tencent (#3) and Alibaba(#13), the two largest Internet companies in China, come into play around 2013 and quickly become very important players in the market.

Similar to results in section 4.2, only communities of bio-ventures (labelled with 'b') and crypto-ventures (labelled with 'c') can be clearly identified. This is a little surprising as we expect a moderate number of communities relate to different investment themes at different times. A possible explanation is that bio-medicine industry requires the most area expertise. As a result, not many investors are eligible to invest in that industry. Internet service, the biggest investment theme in recent decades that contains a large number of industries related to Internet, however, has a much lower barrier to entry. Almost every one in the market invest in Internet related startups. That is why we are not able to identify a community of an industry other than bio-medicine. As for blockchain and cryptocurrencies, they are quite new and not encouraged by Chinese government. Thus, only a specific group of investors invest in relevant projects, forming a community.

6 Discussion

Although our methods above deliver good and interpretable results on Chinese venture capital networks, there are still some limitations in our work.

First, our data is not as complete as we expect. Though it is obtained from a reliable commercial database, the amount is just 1/3 of the estimated entire data. The insufficiency of data shrinks our network size and removes potential edges. That might lead to some inaccuracy or even incorrectness in our analysis.

Second, the way we define our networks makes them inappropriate for some tasks. For example, we can not model a process of information cascade on our networks as the paths do not correspond to real ones that transmit information or substance.

7 Conclusion

In this project, we conduct thorough analysis on networks of Chinese venture capital firms. Our analysis consists of two parts: static analysis and dynamic analysis. For these analyses, we build two types of networks: undirected and directed ones, and investigate both unweighted and weighted undirected networks.

In static analysis, we first find out the network of Chinese venture capital forms a small world. Then we try to extract communities using different community detection methods. Louvain algorithm gives excellent results where most communities have distinct characteristics. Node2vec method can preserve some communities including bio-ventures and crypto-ventures but generate a large community comprised of half the largest venture capital firms which reduces interpretability. We also try to detect communities on weighted networks and the results share the same problem. Another finding is that giving weights to the network can increase modularity but does not generate better communitites.

We also get statistically significant motifs on the directed version of the network. In addition, we show that PageRank can be adopted to the directed network as an alternative effective way of evaluating impact of venture capital firms.

In dynamic analysis, we mainly focus on evolution of communities at different times. We find that the most common pattern of communities in a single year is that it is dominated by one to three large firms, which can be seen as their influence as well. The dominating firms also change over year which is reflective of and consistent with what actually happens in the young but expanding Chinese venture capital industry.

Our project is on Github: `https://github.com/wnls/224w_vc_net`

8 Appendix

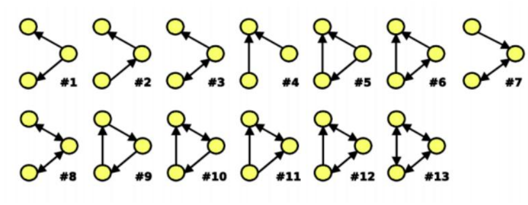

Figure 10: All 13 motifs formed by 3 nodes and directed edges connecting them

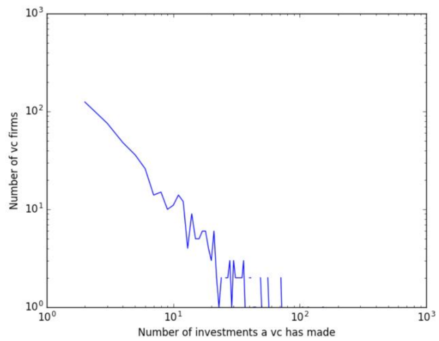

Figure 11: Distribution of the number of investments vc firms have made

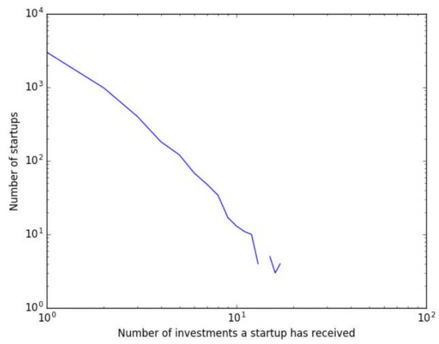

Figure 12: Distribution of the number of investments startups have received

References

[1] Xue, C., Jiang, P. and Dang, X., 2018. The dynamics of network communities and venture capital performance: Evidence from China. Finance Research Letters.

[2] Li, X. and Chen, H., 2013. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. Decision Support Systems, 54(2), pp.880-890.

[3] Liang, Y.E. and Yuan, S.T.D., 2016. Predicting investor funding behavior using crunchbase social network features. Internet Research, 26(1), pp.74-100.

[4] Ng, A.Y., Jordan, M.I. and Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems (pp. 849-856).

[5] Bruna, J. and Li, X., 2017. Community detection with graph neural networks. arXiv preprint arXiv:1705.08415.

[6] Yin, H., Benson, A.R., Leskovec, J. and Gleich, D.F., 2017, August. Local higher-order graph clustering. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 555-564). ACM.

[7] Wu, K., Lee, T. and Ma, A., 2016. Examining the Structure of Venture Capital Investment Networks. CS224W Report. Stanford University. Web.

[8] Hong, Y. and Wu, J, 2016. Linkbase Prediction on Crunchbase Investment Network. CS224W Report. Stanford University. Web.

[9] Shi, J. and Malik, J., 2000. Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), pp.888-905.

[10] Zelnik-Manor, L. and Perona, P., 2005. Self-tuning spectral clustering. In Advances in neural information processing systems (pp. 1601-1608).

[11] Dakiche, N., Tayeb, F.B.S., Slimani, Y. and Benatchba, K., 2018. Tracking community evolution in social networks: A survey. Information Processing and Management.

[12] Jin, Y., Zhang, Q., Shan, L. and Li, S.P., 2015. Characteristics of venture capital network and its correlation with regional economy: evidence from China. PloS one, 10(9), p.e0137172

[13] Jin, Y., Zhang, Q. and Li, S.P., 2016. Topological properties and community detection of venture capital network: Evidence from China. Physica A: Statistical Mechanics and Its Applications, 442, pp.300-311.

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