IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 3, Ser. I (May. –June. 2022), PP 11-14
www.iosrjournals.org

Complete Graph Analysis in Community Detection

Lihong Han
School of Statistics, Lanzhou University of Finance and Economics, Lanzhou, China, 730030
Key Laboratory of Digital Economy and Social Computing Science, Gansu Province, China 730030

Qingguo Zhou
School of Information Science and Engineering
Lanzhou University, Lanzhou, China, 730000

Juheng Zhang
Manning School of Business, University of Massachusetts Lowell
One University Ave, Lowell, MA, USA 01854

Abstract
Community detection in graphs identifies groups and is an essential component of graph theory. The clique percolation method (CPM) has been widely used in graph analysis, but there are computation issues when graphs are large. In this study, we use a Venture Capital dataset from 50 years and show the limitations of the k-clique algorithms. Alternatively, we conducted a complete subgraph search for community detection. The computation time and performance of our complete subgraph search are significantly better than the k-clique algorithm.

Keywords: *network analysis, community detection, k-cliques*

Date of Submission: 06-05-2022 Date of Acceptance: 21-05-2022

I. Introduction

Community detection has been widely used in business, social science, health care, and many other areas. For instance, community detection has been used to identify anti-vax groups on Facebook, dismantle the communities of infected computers when computer virus spreads across the globe, or product recommendations in e-commerce. In the real world, people with common social or economic status live in proximity; in a virtual world, like-minded users frequently interact with each other. Community detection identifies cohesive subgroups in networks and finds communities of nodes that are tightly connected.

The clique percolation method (CPM) has been widely adopted in social network analysis for community detection, along with other methods such as the hierarchical clustering method or classical clustering k-mean algorithms. The CPM starts with cliques with the size k and joins cliques if these k-cliques share k-1 nodes. The method uses cliques as seeds to find communities and it forms communities with the sets of k-cliques being linked and connected with links.

The CPM is computationally challenging in real-world large graphs. We applied the CPM in a Venture Capital (VC) database to identify communities in venture capital investment over 50 years. Although the concept of CPM is straightforward and appealing, the method suffers the issues of scalability and poor community representation. As the k-clique size varies, the computation time can take very long and the performance can get worsen. The CPM is not suitable for large graphs.

In this study, we use a community detection method that is designed for large graphs and apply it to a large venture capital investment dataset. We use complete subgraphs as cliques and find the method is scalable in large networks and provides good performance in identifying communities. In the next section, we will discuss the literature on community detection and venture capital investment. We follow the literature with a discussion on a dataset that we use for this study. Then we discuss our analysis results and conclude the study.

II. Literature

Studies have examined the performance of the CPM and discussed the computation and memory issues of using the method in large graphs (e.g., Baudin 2021; Reid et al. 2012). Reid et al. (2012) discussed the computation issues in the CPM algorithm and proposed a solution to address its performance problems. The

authors also attempted to address its scalability problems. Baudin (2021) proposed a method to improve its memory efficiency.

Existing studies have investigated the impact of VCs syndication networks on the performance of startup companies or ventures (Alexy et al. 2012; Noyes et al. 2014; Ozmel et al. 2013; Yang et al. 2018; Zhang 2018); Zhang and Guler (2019). For example, Alexy et al. (2012) examined the network positions of VCs and found that as the number of connections of a VC firm increased the amount of raised funding increased. Along with degrees, similar results were found with the structural holes of a VC firm and the diversity of a VC firm's syndication network. Hegde and Tumlinson (2014) investigated how social similarities affect VC firms in choosing startups to invest in. The authors found that social proximity, especially ethnicity, was positively associated with startup performance, which was measured by successful exits of startups through IPOs and acquisitions. The authors used two-stage regression and further found that ethnic VCs and startup executives had superior communication and the success of startup performance was largely due to influence through communications among VCs and start-up executives.

The most related work to our study is the work by Yang et al. (2018). Yang et al. used Chinese venture capital firms and studied how network characteristics of venture capital firms affect their performance. The authors used centrality measurements such as degree centrality, eigenvector centrality, and network efficiency to predict a firm's performance. In their study, network inefficiency was defined as the connectedness of a node's neighbors. The more connected an ego VC firm's alter neighbors, the more network inefficiency the ego firm has. The authors found that network inefficiency was a significant positive determinant of a firm's performance. They used the proportion of its funded startup companies that existed through IPOs among all exists (IPOs, merge & acquisition, liquidation, and equity transfer) as the performance measurement. Their study focused on individual VC firms' network position and performance. Other related studies examined how VC syndication network structures affect startup performance. Ter Wal et al. (2016) compared open and closed triads in the VC syndication network and examined how they moderated the impact of knowledge similarity in the syndication network on the success of startups. The authors found that startups were more likely to succeed if their investing syndication was a closed and diverse network or an open and similar network. The success of startups was measured by the probability of receiving second funding. The interaction term of closed/open network and similar/diverse knowledge showed the cofounding effect of the network structure and knowledge sharing.

Other studies have examined the formation of VCs syndication (e.g., Ozmel et al. 2013; Zhang 2018). Zhang (2018) compared novice and serial entrepreneurs and found that when information asymmetry is low (i.e., serial entrepreneurs in the founding team of the startup company), a VC syndicate with higher status, with younger funds, with less familiar members, and greater heterogeneity in type and status are more likely to form. Ozmel et al. (2013) studied the future alliance formation of a startup company with VC firms. The authors found that the effect of the prominence of VCs on a startup company's future funding was significantly positive but diminished as the startup's prominence in its alliance network increased. Mote (2013) used VC firms in the Philadelphia region to study the impact of syndication networks on embracing new types of investments. The author found that syndication was limited but began to emerge in the samples. The VC firms in Philadelphia branched out in the region and co-invested with VC firms outside.

Our study is different from existing studies as we focused on using a different method to identify the co-investment communities in a venture capital dataset. We find cliques and the network position of a clique in the VCs syndication network that affected the performance of startups being invested by the VC firms in a clique. To the best knowledge of the authors, no studies have examined how the connectedness of cliques in the syndication network affects startup success.

III. Data

Data Collection

We used the Venture Capital dataset VentureXpert in U.S.-based ventures and collected all the venture capital firms and startup companies from the year 1962 to 2017. Because companies usually took five years to go on public, we constructed our network of venture capital firms and startup companies from the year of 1962 to 2012. In our sample data, there were 3948 venture capital firms, 96,742 startup companies, and 345,146 venture capital investments. The sampled startup companies were from 18 different industry subgroups.

Conversion of Bipartite Network to Firm Network

To construct our investment network, we project the connections among venture capital firms based on the bipartite network of venture capital firms and startup companies with Pajek. The original bipartite network of venture capital firms and startup companies shows the investment relationship between venture capital firms and startup companies. When a venture capital firm invested in a startup company, they are linked to each other. The original bipartite network has two groups of entities: venture capital firms and startup companies, and the link from a venture capital firm to a startup company shows an investment from the venture capital firm to the

startup company.
The projection converted the original bipartite network into a network of venture capital firms. In our projected network, two firms are linked if they invested in the same company. Our constructed network is unweighted, which means that the frequency of investments is ignored. If two venture capital firms invested in startup companies, they are connected regardless of the amount or the frequency of their investments. We call our constructed network of venture capital firms the firm network.

Community Detection

We conducted a community detection analysis on the firm network by searching for complete graphs with different sizes k. The method we use in this study differs from the popular CPM. When we applied the CPM in our VC network, the computation demanded a large memory and a ten terabytes memory was not enough to complete the computation. Instead of joining overlapping maximal cliques, we use the identified complete subgraphs as our cliques, which vary in different sized. In our dataset, we found that in this network of 3,948 firms there were 2,049 cliques including 1247 isolated VC firms. Isolated VC firms were the firms that hadn't co-invested in a startup company with another firm between the period of the years 1962 and 2017. There were 763 cliques with sizes of three or larger. The largest clique in the network had 45 firms.
The computation steps are summarized in the following Table 1.

Table 1: Community Detection in VC Co-investment Network

Step 1	Input:	The investment activities of VC firms on startup companies.
	Output:	The bipartite network of VC firms and startup companies.
	Process:	Construct the bipartite network of VC firms and startup companies based on the investment. For instance, if a VC firm invested in a startup company, those two nodes (the VC firm and the startup company) form a link.
Step 2:	Input:	The bipartite network of VC firms and startup companies.
	Output:	The co-investment network of VC firms.
	Process:	Covert the bipartite network in step 1 to a network of VC firms based on their co-investment activities. For instance, if two VC firms "A" and "B" invested in the same startup, they form a link.
Step 3:	Input:	The co-investment network of VC firms.
	Output:	The clusters in the co-investment network.
	Process:	Identity cliques in the firm network are based on complete subgraphs. For instance, if VC firms "A", "B", and "C" in the firm network are a maximal complete subgraph in the firm network, they are identified as a clique of size 3.

The cliques we found in the network vary from the size 1 to 45. For each k value, the number of nodes varies from 1 to 1247, the number of venture capital firms.

IV. Empirical Analysis Results

Dependent Variables

We use clusters to find the impact of co-investment on the success of startup companies. We constructed different IPO rates as our dependent variables: IPORate_U, IPORate_I, IPORate_D. The IPORate_U measured the ratio of the IPOs in the union of sets of startup companies that were invested by firms from two cliques. IPORate_I measured the IPOs in the intersection of sets of startup companies that were invested by firms from two cliques. $IPORate\_D$ measured the IPOs in the difference of sets of startup companies that were invested by firms from two cliques.

Independent Variables

Regarding independent variables, we included different groups of measurements, one of which is the network structure features such as various centralities (degree, betweenness, closeness, eigenvector, PageRank), triadic features (clustering coefficients, triangles, triads), ties (strong ties, weak ties).

We conducted regression analysis on different IPO ratios with the static network of firms. Over the sampled period, venture capital firms formed groups as they invested in startup companies. Our first analysis focuses on how cliques or groups in this static network are related to the success of startups to go on public.

We summarized our regression analysis results in Table 2. As shown in Table 2, the clusters in co-investment were significantly positively associated with the performance of startups. The position of cliques in the network, measured by betweenness and eigenvector centrality, was significant. Betweenness centrality measured how often a clique is on a path between other cliques. The significantly negative relationship with IPO rates showed that the position of a clique on the path of other cliques was negatively associated with startup success. In addition, from clique variables, we saw that strong ties were significantly positively associated with

startup performance but weak ties were negatively associated with the IPO rates.

Table 2. Regression Analysis on Static Network Syndication

Variable	IPORate_U	IPORate_I	IPORate0_D
Closeness	0.814768	6.794200	-5.979432
	(11.68925)	(4.553848)	(8.793592)
Betweenness	-0.868451***	-0.320048***	-0.548403***
	(0.059145)	(0.022362)	(0.043117)
Eigenvector	0.06718***	0.0022596***	0.004529***
	(0.01277)	(0.004865)	(0.00939)
PageRank	-4834.711*	-1752.730*	-3081.980*
	(1897.720)	715.8386	(1361.015)
Clustering coefficient	1.465204	-0.098125	1.563329
	(1.475806)	(0.56207)	(1.085372)
Open triads in neighbors	0.001994***	0.000499***	0.001494
	(0.000344)	(0.000131)	(0.000253)
Strong ties	0.878716***	0.441564***	0.437152**
	(0.192717)	(0.073398)	(0.141733)
Weak ties	-0.717713***	-0.233746***	-0.483967***
	(0.082259)	(0.031329)	(0.060497)

V. Conclusion

We searched complete subgraphs to identify communities in the dataset. Compared to the CPM which suffers the memory inefficiency and scalability problems in large social networks, our approach sufficiently identifies communities without causing huge demand on computation time and memory. Our maximal complete subgraphs successfully clustered co-investment relationships among VC firms and predicted the performance of startup companies.

Reference

[1]. Alexy, O. T., Block , J. H., Sandner, P., and Ter Wal, A. L. J. 2012. "Social Capital of Venture Capitalists and Start-up Funding," *Small Business Economics* (39), pp. 835–851.

[2]. Baudin, A. 2021. "Clique Percolation Method: Memory Efficient Almost Exact Communities."

[3]. Hegde, D., and Tumlinson, J. 2014. "Does Social Proximity Enhance Business Partnerships? Theory and Evidence from Ethnicity's Role in U.S. Venture Capital," *Management Science* (60:9), pp. 2355-2380.

[4]. Mote, J. 2013. "Syndication, Networks and the Growth of Venture Capital in Philadelphia, 1980–99," *Industry and Innovation* (18:1), pp. 131–150.

[5]. Noyes, E., Brush, C., Hatten, K., and Smith-Doerr, L. 2014. "Firm Network Position and Corporate Venture Capital Investment," *Journal of Small Business Management* (52:4), pp. 713-731.

[6]. Ozmel, U., Reuer, J. J., and Gulati, R. 2013. "Signals across Multiple Networks: How Venture Capital and Alliance Networks Affect Interorganizational Collaboration," *Academy of Management Journal* (56:3), pp. 852-866.

[7]. Reid, F., McDaid, A., and Hurley, N. 2012. "Percolation Computation in Complex Networks."

[8]. Ter Wal, A. L., Alexy, O., Block, J., and Sandner, P. G. 2016. "The Best of Both Worlds: The Benefits of Open-Specialized and Closed-Diverse Syndication Networks for New Ventures' Success," *Adm Sci Q* (61:3), pp. 393-432.

[9]. Yang, S., Li, Y., and Wang, X. 2018. "Cohesiveness or Competitiveness: Venture Capital Syndication Networks and Firms' Performance in China," *Journal of Business Research* (91), pp. 295-303.

[10]. Zhang, L. 2018. "Founders Matter! Serial Entrepreneurs and Venture Capital Syndicate Formation," *Entrepreneurship Theory and Practice*), pp. 1–25.

[11]. Zhang, L., and Guler, I. 2019. "How to Join the Club: Patterns of Embeddedness and the Addition of New Members to Interorganizational Collaborations," *Administrative Science Quarterly*).

Lihong Han, et. al. "Complete Graph Analysis in Community Detection." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(3), 2022, pp. 11-14.