JOIM
www.joim.com

JOURNAL OF INVESTMENT MANAGEMENT, Vol. 17, No. 1, (2019), pp. 92–103

SURVEYS AND CROSSOVER

This section provides surveys of the literature in investment management or short papers exemplifying advances in finance that arise from the confluence with other fields. This section acknowledges current trends in technology, and the cross-disciplinary nature of the investment management business, while directing the reader to interesting and important recent work.

PREDICTING INVESTOR SUCCESS USING GRAPH THEORY AND MACHINE LEARNING*

Jeffrey Glupker[a,b], *Vinit Nair*[a,c], *Benjamin Richman*[a,d], *Kyle Riener*[a,e] *and Amrita Sharma*[a,f]

We extract a large dataset of venture capital financing and related startup firms from Crunchbase. This paper examines how network position determines the success rate of investors. Precision in determining which investors will be successful is relatively high, but it is in fact easier to predict unsuccessful investors. Graph-theoretic features may be used in machine-learning algorithms to improve predictions of VC performance. This study has implications for how startups and private bank investors may choose investors and suggests a two-step approach where segmentation by industry is done first, followed by community construction within industry. In short, choosing a VC should be first based on subsetting VCs who have a focus in the industry of the startup followed by the use of a machine-learning model. This cross-disciplinary paper generates insights by combining financial data with graph-theoretic ideas and machine-learning algorithms.

*The authors would like to thank the technical staff at Credit Suisse Labs for their mentoring in this project.
[a]Santa Clara University, Santa Clara, CA 95053, USA.
[b]E-mail: jglupker@scu.edu
[c]E-mail: vnair@scu.edu
[d]E-mail: brichman@scu.edu
[e]E-mail: kriener@scu.edu
[f]E-mail: asharma2@scu.edu

1 Introduction

The success or failure of a startup company is influenced by many factors, but the fact remains that these young firms need funding to get their businesses off the ground and continue operating. We believe that, in addition to typical success predictors such as funding amount, who finances

these startups is a key factor to be considered. If this were not true, all investors would been seen as equal: an investment of $500k from Investor A would be exactly the same as $500k from Investor B.

Aside from an investor's business knowledge and acumen—both of which can be difficult to measure accurately and objectively—we believe that their positioning within the larger investor network is an influential component of their investment success or failure. Within the entire network, we anticipate finding communities of investors that are differentiated by which industries they invest in and the funding stages in which they commonly participate. It is likely that these communities have varying levels of success.

From the perspective of the startup company, having the ability to identify the most influential and successful investors in the communities most relevant to their mission has obvious allure. Our primary goals are to determine the overall connectedness among a network of investors; examine various communities of investors within the network and highlight their differentiating characteristics; and calculate how influential an investor's position within that network is in determining whether that investor is successful.

Our journey to answer these questions combines the financial elements of venture capital investment with advanced quantitative methods of graph theory, which are used to test our hypotheses. By marrying these two disciplines—finance and computer science—we hope to contribute to the literature in both fields and gain insights not possible to find in one without the other. Using Crunchbase as our source, we gathered information on startup investment activity from the 1900s up to 2013, focusing our analysis on the most recent 10 years. With the industry evolving as quickly as it has, our belief is that the most recent data will bring the most relevance to understanding any future implications.

We first study a network of investors who are connected by edges representing common investments. These edges are weighted based on the number of shared investments between two investors; these shared investments must have occurred during the same funding round and time period. Once the network is established, we calculate connectedness measures such as degree and clustering coefficients. These network metrics are inputs into a model that attempts to predict how likely an investor is to be successful. Defining exactly what success means is not a black and white problem, and can take on many values. Our measurement methodology defines success as the share of an investor's total companies invested in that have exited through IPO or acquisition. Given this definition for success, we find that 17.95% of all investments made within our data have been successful. While we understand that additional factors may be used in the definition of success, we utilize this as a starting point. Other considerations include the number of funding rounds, the most recent stage, time between rounds, and total amount raised.

Section 2 briefly reviews some related research on the topic. Section 3 follows by going into detail regarding the data used to conduct this analysis, including key definitions of metrics, dimensions, and scope. We then dive into our analytical process, reviewing the specifications of network construction and individual community analyses, and follow that with our predictive models for success (Sections 4 and 5). We conclude in Section 6 with a review of our machine-learning models examining the contribution of network metrics in investor success, along with some insights and potential applications.

2 Previous research

With the objective of our research stated as predicting the success of an investor using network position, our scope is primarily focused on which metrics of the network graph could have a significant influence. Our study does not take into account how an investor's demographics or social characteristics influence their network position.

Gupta *et al.* (2015) illustrate a method for identifying successful investors by analyzing how an investor's collaboration network changes over time. The identified methodology, InvestorRank, is used to capture the intuition of becoming close to an exemplary investor or super-angel over time. The result of the research shows potential in discovering investors who will become successful. InvestorRank provides two heuristics to identify successful investors:

(1) Flag investors whose InvestorRank is consistently below 100.
(2) Flag investors who follow a general trend of improvement when compared to their preceding snapshot based on a threshold.

Based on these two heuristics, of the 1,524 unlabeled investors the second rule flags potential rising stars effectively, such as General Catalyst Partners and Paul Buchheit.

However, the premise of Gupta *et al.* (2015) does not take into account the investment amount, number of investments, funding rounds, time of funding, or entity type (individual or firm) of the investor. For instance, the dataset takes all of the investments into account, but 55% of investors made just one investment. Considering these observations in the analysis adds a high amount of uncertainty, as an investor cannot be deemed successful based on a single investment.

The exemplar list being considered as a reference point to measure the closeness of an investor was formulated from Wikipedia and other online articles. Conventional methodology states that the success of an investor could be defined based on whether a company invested in went public, or was acquired or merged with another firm. However, there was no clear definition of success used in Gupta *et al.* (2015) by which an investor is assessed. The paper also does not consider certain major events such as Y2K, the dotcom bust, and the 2008 recession which could have inflicted a major impact on the startup ecosystem.

The common theme between InvestorRank and our study is that the startup network has a significant influence on investor success. The methodology used by InvestorRank aims to determine which unknown investors are successful, and thus will be useful for companies looking for funding.

Bubna *et al.* (2018) performed research that ours closely parallels, defining venture capital communities and researching the effect of funding from a community VC on the likelihood of a successful exit. Multiple definitions were used for a successful exit, with one including both IPOs and M&As. Bubna *et al.* (2018) also used the same community detection methodology as us, but included variables such as VC age and functional style. The key difference from our paper is that the dependent variable is the success of the startup. There is also no distinction between VC communities; all VCs that belong to a community, no matter the size, are labeled "community VC".

Sorenson and Stuart (2001) also investigate VC communities but focus on their geographical and social distance from entrepreneurs seeking funding. Their paper then explores the impact of these different forms of distance on the likelihood of the parties engaging in exchanges. Success of the business is not considered.

Adcock *et al.* (2012) have done similar work, analyzing the bipartite investment network and making a comparison of this network to previously studied social and information networks, with link prediction serving as the end goal. Their investment network has investors and investees on the nodes and investment in the company as the edge. The paper flattens the network to identify interesting network structures for different investor types. For instance, personal investors have the highest average clustering coefficient, which indicates that they tend to group together. And tech companies have the lowest average clustering coefficient which indicates that they choose to acquire small firms rather than invest in them.

Adcock *et al.* (2012) modified a weighted ranking algorithm, applying it to a bipartite investment graph along with rule-based filtering. The weights in the algorithm are defined using Common Friends, Jaccard Distance, Adamic-Adar, or Preferential Attachment weighting methods. The combination of modified Jaccard weights and rooted PageRank gives the best results.

A limitation of Adcock *et al.* (2012) is its low precision in link prediction compared to the social networks due to lower edge count. The algorithm also did not consider repeated investments; however, a significant portion of test edges were repeated edges. A commonality between our paper and Adcock *et al.* (2012) is using network metrics to determine the success of an investor; they also included these for link prediction. However, we have also used machine learning along with network metrics in order to classify investor success. Our paper has attempted to handle the majority of the extreme scenarios such as repeated investments between two investors, single investments and category-wise investments, making our data model quite robust.

In contrast to existing papers, our focus is to predict investor success using network positioning, first by defining the success of an investor, and second by considering the key network metrics in our model. The hypothesis in our work is that investors that are very well connected may have a significant impact on the entire start-up ecosystem, and in turn see greater success.

3 Data

The data for our research is sourced from Crunchbase, which, according to their website, "was founded to be the master record of data on the world's most innovative companies".[1] Included are characteristics of worldwide startup companies, investors, and individuals throughout the 20th century and up through 2013.

There are over 17,500 startup companies in our dataset. Additional background information about each company such as their industry, founding date, amount of money raised and number of investors per funding round is also included. One of the more important variables is company status, which informs whether a company has successfully exited (via IPO or acquisition), has closed, or is still operating.

From the entire dataset, we have narrowed our focus to activity that occurred within the final 10 years (2004–2013) and limited these to only companies founded in the US. The reason for this cutoff is due to the large increase in activity during the mid-2000s, as shown in Figure 1. A few key factors figured prominently in this boom. According to Rose and Grof (2016), investment activity was spurned in 2005 after Y Combinator introduced the accelerator model to facilitate start-up growth. Also around this time, the technology crucial to operating a new business was becoming cheaper and easily accessible (L.S., 2014).

Due to over 53% of investors having made just a single investment, we further restricted our

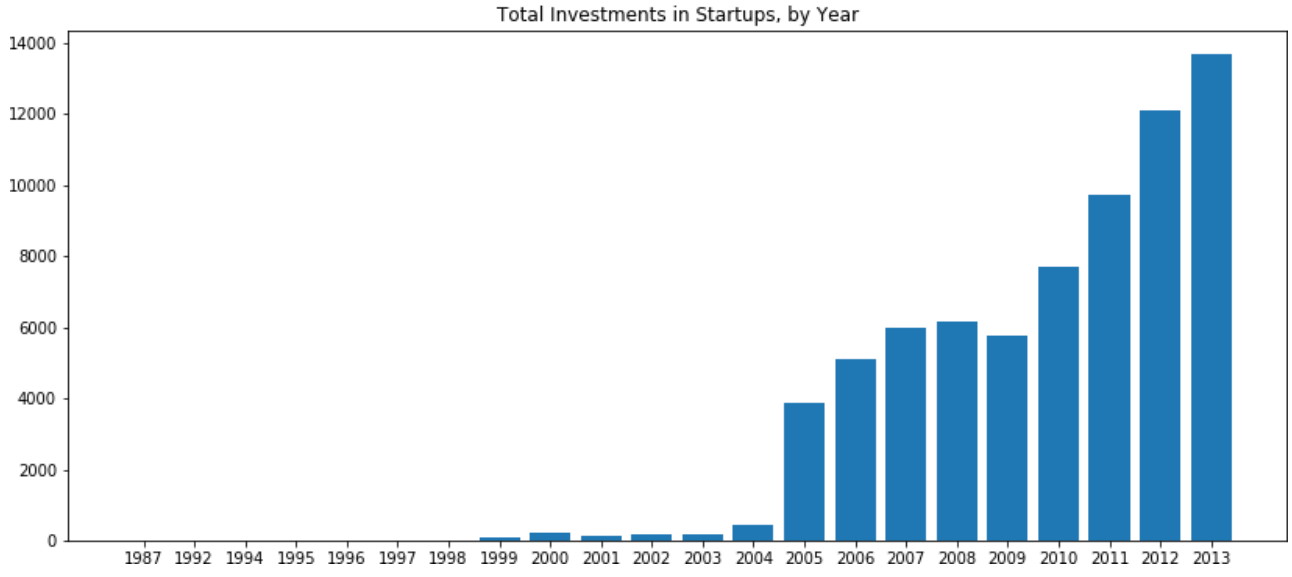

Figure 1 Investments by year reported on Crunchbase. We see a boom of activity starting in 2005 with the introduction of accelerator models followed by stagnation from 2007 to 2009 during the Great Recession.

analysis to only those investors who have made multiple investments. Our final dataset includes over 14,000 startups and over 5,300 investors, for a total of nearly 60,000 investment rounds.

3.1 Key metrics and dimensions

The dataset used is a combination of multiple individual files and contains fields pertaining to individual investments broken down by funding rounds and investment dates.

Apart from the given fields in the dataset, a number of other variables were calculated and used during the course of this analysis. These include variables that convey how connected investors are (centrality) and position of investors within a network (community). A complete list of important variables and their definitions is provided in the Appendix.

3.2 Data engineering

Crunchbase categorizes startup companies into one of forty-two industries, ranging from nano-technology to advertising, music to mobile, and many others in between. We identified similar industries and narrowed the list down to the seven groups below to help aid in our analysis:

(1) Business: includes major industries like automotive, education, and manufacturing
(2) Entertainment: includes categories like sports, video games, and fashion
(3) Health: includes categories listed as medical, health, and biotech
(4) Services: includes categories such as advertising, legal, and transportation
(5) Tech – General: includes categories such as software, messaging, and network hosting
(6) Tech – Hardware: includes categories such as semiconductors and nano-technology
(7) Other: all categories that did not fall into any of the above.

There are two reasons for doing this. First, for a new firm seeking out investors it makes sense that they would look to investors who have had experience investing in companies similar to their

own. Second, as certain industries have seen more rapid growth than others (like Tech), we isolate them so as not to skew the overall results.

3.3 Creating the network

The dataset is flattened out into a network where investors are nodes and edges represent common investments. These edges are weighted based on the number of shared investments between two investors; these shared investments must have occurred during the same funding round and date. For each of these nodes, graph-theoretic metrics like degree centrality, clustering coefficient, and eigenvector centrality are calculated. Within our network, we have nearly 5,300 total investors, with an average degree (number of links) of 23.79. The overall connectedness of the network is relatively high, a likely result of only examining investors with multiple investments, as just over 6% of investors have only one degree and less than 13% have two or fewer. Finally, using Louvain heuristics (Blondel *et al.*, 2008) nodes are grouped into communities, more details of which are given below.

In addition to creating the overall network, we also constructed networks specific to investors within the following industries: Tech-General, Tech-Hardware, and Health. The remaining four categories were grouped into a single network and further analyzed.

3.4 Community detection

For our analysis, we use a package in the Python programming language called "python-louvain" for community detection. The Louvain method (Blondel *et al.*, 2008) detects communities by optimizing modularity. The modularity of a partition of a network is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities. In case of weighted networks, such as ours, it is defined as

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{K_i K_j}{2m} \right] \delta(c_i c_j)$$

where,

- A_{ij} represents the weight of the edge between i and j
- K_i is $\sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i
- c_i is the community to which vertex i is assigned
- $\delta(c_i, c_j)$ is 1 if i and j belong to the same community, i.e., if $c_i = c_j$, and 0 otherwise
- $m = \sum_{ij} A_{ij}$ is the total weight of edges in the network.

The Louvain method attempts to optimize the modularity of a partition in two steps,

(1) *Modularity optimization* – Assigning and reassigning nodes to communities by maximizing modularity
(2) *Community aggregation* – Communities are aggregated (treated as nodes) to build new net work of communities.

The above steps are repeated iteratively until modularity can no longer be maximized further. The Louvain method is also preferable because it can detect communities in very large networks relatively fast.

Within the network, we have identified 79 independent communities of investors that range widely in regards to size and diversity of investments. Only two communities include over 1,000 investors (see Table 1), while a large majority (86%) of these communities are very small and include fewer than 10 investors. A breakdown of the top-10 by volume is shown below; not surprisingly, 91 of the top-100 investors by the centrality measure are part of one of the first two communities.

Table 1 Top-10 communities by size. We see many investors grouping together, as these 10 communities account for 95% of all investors in our network.

Community rank	Community size	Cumulative size	Cumulative %
1	1,251	1,251	24%
2	1,243	2,494	47%
3	833	3,327	63%
4	432	3,759	71%
5	270	4,029	76%
6	267	4,296	81%
7	243	4,539	86%
8	230	4,769	90%
9	222	4,991	94%
10	71	5,062	95%

4 Research objective and analysis

Our analysis serves multiple purposes. The first is to determine how accurately we can predict which investors will be successful or not based on their positioning within an investor network, in addition to characteristics related to their investment practices. This serves to aid entrepreneurs in determining who to pitch to and partner with, as they are looking to connect with the investors who give their startup the best chance for success.

The second is to quantify the benefit of examining investor success based on smaller, niche networks rather than the broader network that includes all investors. We hypothesize that an investor's positioning within a smaller network of peers will result in more accurate results. Say there are two startups, one that has discovered a better way to test blood samples, and another that has a platform that allows for seamless peer-to-peer payments. While looking at the entire network of investors to find the best potential partners may pan out, we believe that it would be better for the first startup to examine the network VCs who have invested in health-related companies, and the second startup to examine the network of VCs who have a history of investing in FinTech.

We split the data into two, five-year periods (2004–2008, 2009–2013), using the first five years to train the model and the second five years to test on. By doing this, the results are intended to show how well current performance can predict future success.

To proceed with our analyses, we classified investor success as a binary outcome: if at least 20% of the startups a given VC backed financially have exited successfully, they are classified as successful. If the VC falls below that threshold, they are not.

We also constructed individual networks for the health and technology industries listed above using the method used in Section 3, with the remaining groups (business, entertainment, services, others) aggregated into a single network. Only investors who had invested within these industries were considered, however they did not have to be exclusive investors within the industry (i.e., they could also have investments in any of the other categories).

4.1 Community categorization

Two networks of communities were created for each industry group, one for each of the five-year periods (2004–2008 and 2009–2013). While the number and size of communities varied with each network, our goal was to create a standard framework across each network to determine which communities should be classified into a common set of subgroups.

This involves a two-step process. Looking at the individual community distributions within a network, some appeared to lend themselves more to having three distinct community groupings (small, medium, large) while others were more conducive to just two groupings (small, large).

Our first step was to determine whether two or three community groupings would best fit the network; the next step was to determine exactly how communities would be classified into these groupings.

To accomplish this, we first selected initial community size cutoffs for both two and three community categories based on the community distribution within each network. For two community categories, there are two cutoff values to optimize: a cutoff between small and large community groups for each five-year window of data, $\vec{c} = [c_1, c_2]$ where c_1 is the cutoff for 2004–2008 and c_2 for 2009–2013. For three community categories, there were four cutoff values to optimize: cutoffs between small–medium, and medium–large communities for each five-year window of data, $\vec{c} = [c_1, c_2, c_3, c_4]$ where c_1 and c_2 are the low and high cutoffs for the first time period and c_3 and c_4 are for the second time period.

A strategy was then implemented that would begin with an initial cutoff value guess, $\vec{c}$, and methodically search a grid of cutoff values around that starting point. We created an initial set of steps for each cutoff to take, such that for each cutoff value c_i and corresponding step length s_i, the set of eligible guesses are $\{c_i - s_i, c_i, c_i + s_i\}$. For small-large community cutoffs, this resulted in 9 total possible cutoff combinations to search across, and for small–medium–large community cutoffs this creates 81 total possible cutoff combinations.

Each cutoff combination categorizes a particular community as being within a grouped community size. The grouped data is run through a series of classification algorithms, returning a mean precision score. There will be 9 precision scores for the small–large community combinations, and 81 for the small–medium–large combinations. The single set of cutoff values within each of these which register the highest mean precision scores are returned.

Starting with two once-optimized cutoffs as the new initial guess, new cutoff grid searches are formed around these using smaller steps. The grid search optimization is performed the same number of times between the small–large community

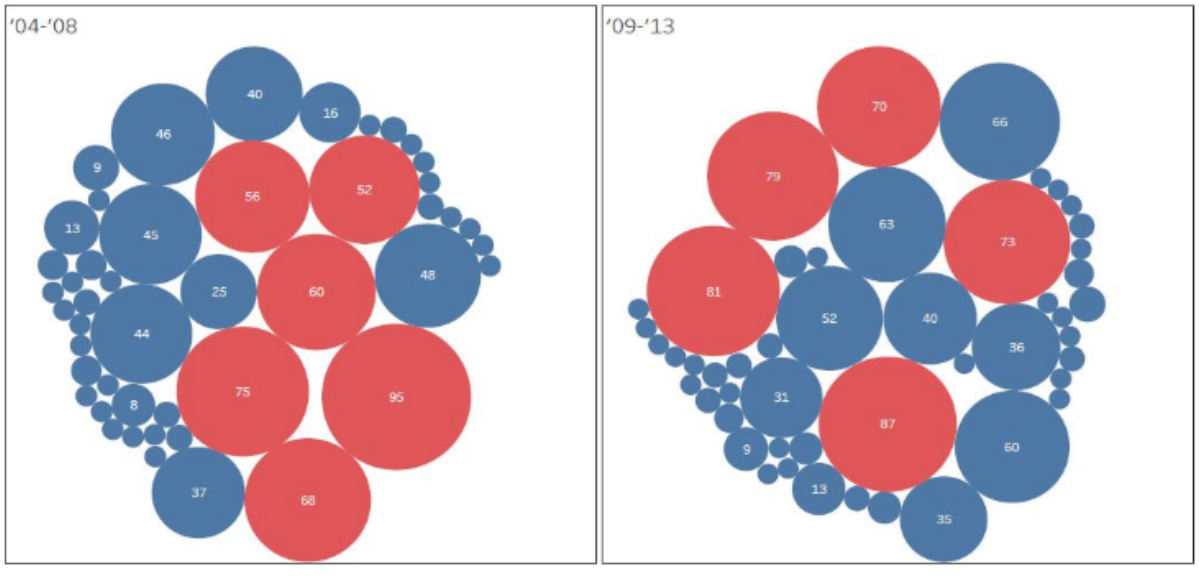

Figure 2 Communities within the Tech-Hardware network for each five-year period are shown above. Based on our community detection algorithm, those shaded in red were classified as large communities.

cutoffs and the small–medium–large community cutoffs. After the grid-search optimization is completed, the two returned cutoffs are the most-optimized cutoff set for the small–large division and the small–medium–large division. The highest mean precision score between the two determines both the optimal set of cutoff values, and whether the particular industry is better modeled as having two or three distinct community sizes.

This process is repeated for a set number of epochs, depending on the size of the difference between communities. The greater the difference in the community sizes, the fewer the epochs needed to determine optimal cutoffs. An example of the final classification for the Tech-Hardware network can be seen in Figure 2.

Since our final results come from an ensemble learning method across models, we wanted to ensure that the cutoffs were optimized simultaneously for all models included in the ensemble method. To do this we ran chose the cutoff combination which produced the highest average precision results across all models.

4.2 Ensemble learning

After community categorization had been determined for all networks, we then moved forward by leveraging an ensemble learning method to test our hypotheses. For each network, training was done on the 2004–2008 data. In many models the distribution of the target variable outcomes was evenly split, however for instances where this was not the case we rebalanced using over-sampling techniques to avoid any uneven skew. An odd number of models were chosen in the ensemble to ensure majority rule when voting: across all predictors, whichever was the most common predictor was selected and compared with the actual results.

Because our target variable is binary, models used in the ensemble include both traditional and classification models, such as Logistic Regression, Random Forest and K-Neighbors.

Since we are most interested in understanding how often the VCs that we predict to be either successful or unsuccessful actually are, we look to precision our primary measurement. In addition to computing precision results for each class, accuracy, recall, and F_1 scores were also considered. Recall is defined as the ratio of the number of correctly predicted instances of a class to the total number of true instances of that class, i.e., of the total number of observations of a class, the percentage that we identified. The F_1 score combines these metrics by taking their hyperbolic mean: $F_1 = \frac{2}{\frac{1}{Recall}+1\,Precision}$.

5 Results

Our analysis produced varying results across each of the networks we studied, as shown in Table 2. While the overall accuracy ranges between 55 and 75% based on the industry, we find that the largest, overall network has a low Type 1 error rate (false positives), yet the highest Type 2 error rate (false negatives). The accuracy of this network is near the weighted average of the individual networks, and is brought down somewhat by the low accuracy of the Tech-General network. Low accuracy here could be for any number of reasons, and with

Table 2 Accuracy and error rate, by industry category.

Industry category	Accuracy	Type I error rate	Type II error rate
Overall network	64.9%	25%	61%
Tech-General	54.7%	50%	32%
Tech-Hardware	69.6%	31%	32%
Health	73.0%	21%	47%
Other industries	63.2%	35%	39%

the growth in this industry during the given time window, subsetting this even further may produce higher accuracy.

As mentioned in Section 4.2, precision is our primary metric over accuracy because we are looking to understand how true our individual predictions of successful and unsuccessful investors are. Here is where we find more insightful results: although our hypothesis was rooted in a theory of predicting successful investors, what we have found is that we are much more likely to identify the unsuccessful investors. While unexpected, the application of this learning can still be relevant to our theory. From the perspective of startup founders, knowing which VCs are harbingers of success and which would likely result in failure are two sides of the same coin. And because the funding process is ultimately the decision of the investor, not the entrepreneur, there is no guarantee that a startup will get their first choice of VC to partner with. Knowing that the results vary by industry, having the ability to group VCs into those deemed unsuccessful and pursuing the rest (or vice versa) still allows for some paring of the entire investor ecosystem into only those that would seemingly provide the best path forward.

Table 3 further highlights the difference by industry, most notably that for each network the precision of identifying the unsuccessful investors is significantly higher than that of identifying the successful ones. Each network studied produced precision results for the unsuccessful investors between 75 and 80%, with the smaller networks having marginally better numbers than the overall network.

Due to the challenging nature of startup investment, it may come as little surprise that the results vary by industry. However, the fact that subsetting the network into smaller graphs produces results that vary from the overall network would imply that industries do behave differently, and there is value in more detailed analytical approaches.

Table 3 Classifying successful vs unsuccessful investors.

	Precision	
Industry category	Successful investors	Unsuccessful investors
Overall network	38%	75%
Tech-General	40%	76%
Tech-Hardware	58%	80%
Health	56%	77%
Other industries	46%	77%

6 Conclusion and applications

Being able to narrow down the choice of investors has many useful applications. For a private bank, value would come from recommending investors to clients interested in launching their own startup. Knowing which investors are less likely to succeed provides a significant advantage. Startups could save time and effort when they are informed who to focus their pitches on. Given the busy lifestyle of founders, it is crucial to save every minute possible in order to run the business. This can be just as important as the funding itself in reaching the ultimate goal of success. Knowing which investors to go to or avoid would also give clients the advantage of being first movers. While competitors may struggle to engage with as many investors as they can, clients can go directly to the successful ones and present their product first. The fortune of a startup is unpredictable, so direction provided by the model would be a desirable benefit for customers as it could relieve some of the worry and position the client better than their competitors.

If any clients have interest investing in startups, this model can suggest how to improve their chances of success. The most important factor

we found was being connected to many other investors. A bank can help expand a client's network by using its own network to introduce them to fellow investors. We also explored how different industries affect investor success, which can help a novice investor who is more interested in getting his feet wet than devoting himself to a particular industry. We saw mixed results between industries, but there are some specific job sectors that perform better than the rest and are worth keeping an eye on.

Armed with this analysis, a bank can provide crucial intelligence to its customers. Investing in startups is a struggle for all parties involved and any little advantage can be the difference between success and failure. Our work aims to help eliminate some of the uncertainties of the field for the benefit of those in our own network.

Appendix: Variable definitions

Variable	Description
Investor name	Name of the investor.
Investor entity type	Type of investor (Company, Financial Organization, Individual investor).
Investee name	Name of the company invested in.
Investee category	Category of the company invested in (Biotech, Finance, Software, etc.).
Investor category	Category of the investor (Biotech, Finance, Software, etc.). Only applicable if the investor is a company.
Investment round	The funding round that the investment was made in (crowd funding, private equity, venture, etc.).
Date of investment	Date of the investment.
Amount raised per funding round	Total amount raised in a specific funding round and date by all participants combined.
Investee current status	Current status of the company invested in (operating, closed, etc.).
Number of investors	Number of investors per funding round and date.
Success ratio	Proportion of successful exits out of all investments made by an investor.
Eigenvector centrality	Measure of connectedness and influence of an investor.
Community id	Community that an investor is a part of.
Majorcommunity/mid-sized community	Binary variables that indicate if an investor is a part of a large community or a mid-sized community.

Note

[1] See: https://about.crunchbase.com/about-us/.

References

Adcock, A. B., Lakkam, M., and Meyer, J. (2012). "Cs 224w Final Report Group 37," Term Paper.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment 2008* (10), P10008.

Bubna, A., Das, S. R., and Prabhala, N. (2018). "Venture Capital Communities," *Journal of Financial and Quantitative Analysis*, forthcoming.

Gupta, S., Pienta, R., Tamersoy, A., Chau, D. H., and Basole, R. C. (2015). "Identifying Successful Investors in the Startup Ecosystem," in *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, New York, NY, USA, pp. 39–40. ACM.

L. S. (2014). "The Startup Explosion," *The Economist.*

Rose, D. S. and Grof, M. (2016, June). "The State of the Startup Accelerator Industry," *Forbes.*

Sorenson, O. and Stuart, T. (2001). "Syndication Networks and the Spatial Distribution of Venture Capital Investments," *American Journal of Sociology*, **106**(6), 1546–1588.

Keywords: Venture capital; entrepreneurship; investor success; social network analysis; community detection; graph theory; machine learning