Venture Capital Network Structure and Nestedness Analysis

Disclaimer: this is a intermediary report over the results gotten until now. Colored rectangles represent "to do" or "ongoing" activities to be done in subsequent work until the final version of this article is complete.

1 Methodology

1.1 Data Source and Preprocessing

This study uses data from Crunchbase, a broad database containing information about startups, venture capital firms, and investment rounds. The dataset includes information about companies, investors, investments, and funding rounds in the United States market. International venture capital firms from other countries also appear in the dataset when they participate in US startup investments.

The data preprocessing follows established methodologies from entrepreneurship literature [3]. The cleaning process implemented includes several steps: (1) removal of companies with incomplete information, (2) exclusion of companies founded after 2017 to allow sufficient time for investment patterns to emerge, (3) removal of companies with exit status (bankruptcy, acquisition, or IPO), and (4) application of a minimum funding threshold of $150,000 to focus on substantive investment relationships.

1.2 Investment Network Construction

The analysis focuses on venture capital co-investment patterns across different funding stages. Investment stages are categorized into two main groups:

- Early stages: angel, pre-seed, seed, and Series A
- Late stages: Series B through Series I

A bipartite network is constructed where nodes represent venture capital firms and edges represent co-investment relationships in the same company. The network is bipartite because it connects two distinct sets of investors: those participating in early-stage rounds (right nodes) and those participating in late-stage rounds (left nodes).

This approach allows us to study how early-stage and late-stage investors interact in the investment ecosystem.

The bipartite graph $G = (U \cup V, E)$ consists of:

$$U = \{u_1, u_2, \ldots, u_m\} \text{ (late-stage VCs)} \tag{1}$$

$$V = \{v_1, v_2, \ldots, v_n\} \text{ (early-stage VCs)} \tag{2}$$

$$E \subseteq U \times V \text{ (co-investment relationships)} \tag{3}$$

To prevent spurious connections from related entities, investor pairs where the first five characters of their names match are filtered out, reducing the likelihood of including different funds from the same parent organization. Further more, investors that participated in both early and late stages receive a suffix so they can be threated as distinct agents for each phase.

Clearly show the overlap or number of connections made between the same investors but in distinct phases ex. VC1_serieA-VC1_serieC

1.3 Community Detection

Community structure in the bipartite network is identified using the greedy modularity optimization algorithm [2]. This method iteratively merges communities to maximize the modularity score, which measures the density of connections within communities compared to connections between communities.

For a bipartite network, modularity Q is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{4}$$

where A_{ij} is the adjacency matrix, k_i is the degree of node i, m is the total number of edges, c_i is the community of node i, and $\delta(c_i, c_j)$ is 1 if nodes i and j are in the same community, 0 otherwise.

The algorithm identifies communities of venture capital firms that frequently co-invest together, revealing structural patterns in the investment ecosystem that may not be apparent from individual investment decisions.

1.4 Nestedness Analysis

Nestedness is a structural property commonly observed in ecological networks [1] that describes the tendency for specialists to interact with a subset of the partners of generalists. In the context of venture capital networks, nestedness would indicate that investors with fewer connections tend to co-invest with a subset of the partners of more connected investors.

We measure nestedness using the NODF (Nestedness based on Overlap and Decreasing Fill) metric [1]. For a bipartite adjacency matrix M with rows and columns sorted by decreasing degree, NODF is calculated as:

$$NODF = \frac{NODF_{rows} + NODF_{columns}}{2} \tag{5}$$

where:

$$NODF_{rows} = \frac{100}{R(R-1)/2} \sum_{i=1}^{R-1} \sum_{j=i+1}^{R} \frac{|N_i \cap N_j|}{k_j} \text{ if } k_i > k_j \tag{6}$$

$$NODF_{columns} = \frac{100}{C(C-1)/2} \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} \frac{|N_i \cap N_j|}{k_j} \text{ if } k_i > k_j \tag{7}$$

Here, R and C are the number of rows and columns, N_i represents the set of connections for node i, and k_i is the degree of node i.

With this method, NODF vary between 0 and 1 (perfect nestedness).

1.5 Statistical Significance Testing

To determine whether observed nestedness values are significantly higher than expected by chance, we employ a null model approach using the Curveball algorithm [4]. This algorithm generates randomized matrices that preserve the degree sequence of both node sets while randomizing the connection patterns.

For each community, we generate 100 null matrices using 10,000 Curveball iterations. The statistical significance is assessed by comparing the observed NODF score against the distribution of null model scores:

Generate 1000 null matrices instead

$$Z = \frac{NODF_{observed} - \mu_{null}}{\sigma_{null}} \tag{8}$$

where μ_{null} and σ_{null} are the mean and standard deviation of the null distribution. Communities with $p < 0.05$ (where p is the proportion of null models with NODF $\geq$ observed NODF) are considered to have significantly high nestedness.

Better explain Z-core and P-values interpretation and relationships

2 Results

2.1 Network Characteristics

The Crunchbase dataset, following the cleaning processes described in the "Methodology" section, yields 147,832 investment registers, representing transactions among 22,527 companies and 38,843 investors.

Exclusion of non-venture capital investors reduces the dataset to 104,618 investment records and 16,932 unique companies with venture capital funding.

The division of venture capital firms into early-stage and late-stage investor groups results in 169,679 investment pairs comprising 3,666 unique startups.

Add network visualization showing bipartite structure

2.2 Community Structure and Size Distribution

Community detection using greedy modularity optimization identifies 175 distinct communities, with the largest communities containing over 4000 investors pairs each.

Analysis focuses on communities with at least 150 nodes to ensure statistical power for nestedness analysis. This threshold excludes smaller communities that may not provide reliable nestedness measurements due to limited connectivity patterns. Such a threshold yields 5 communities.

Table 1 shows the size distribution of the largest communities identified by the modularity optimization algorithm.

Rationale of threshold

The largest three communities (0, 1, and 2) contain over 12,000 investors combined, representing approximately 75% of all investors in the network. This concentration suggests a highly centralized structure within the venture capital ecosystem, with most investment activity occurring within a small

Community ID	Number of Pairs
0	4,248
1	4,089
2	3,959
3	979
4	188
5	155
6	137
7	122

Table 1: Size distribution of the largest investor communities identified through greedy modularity optimization

number of large communities.

Mention literature, as this phenomena is somehow well-known

The community size distribution follows a typical power-law pattern observed in many social networks, where the top 5 communities by size account for the majority of investors in the network, suggesting a hierarchical organization within the venture capital ecosystem.

Add figure of community size distribution

2.3 Nestedness Findings

Nestedness analysis across investor communities reveals heterogeneous structural patterns. Among the 5 communities examined, one exhibits statistically significant nestedness ($p < 0.05$) relative to degree-preserving null models generated through the Curveball algorithm.

Figure 1 presents the comparison between observed and null model nestedness scores, where each data point represents a distinct community positioned according to its observed NODF value against the corresponding null model mean.

Community 2 demonstrates the most pronounced nestedness, exhibiting an NODF score of 0.088 with statistical significance of $p = 0.00001$.

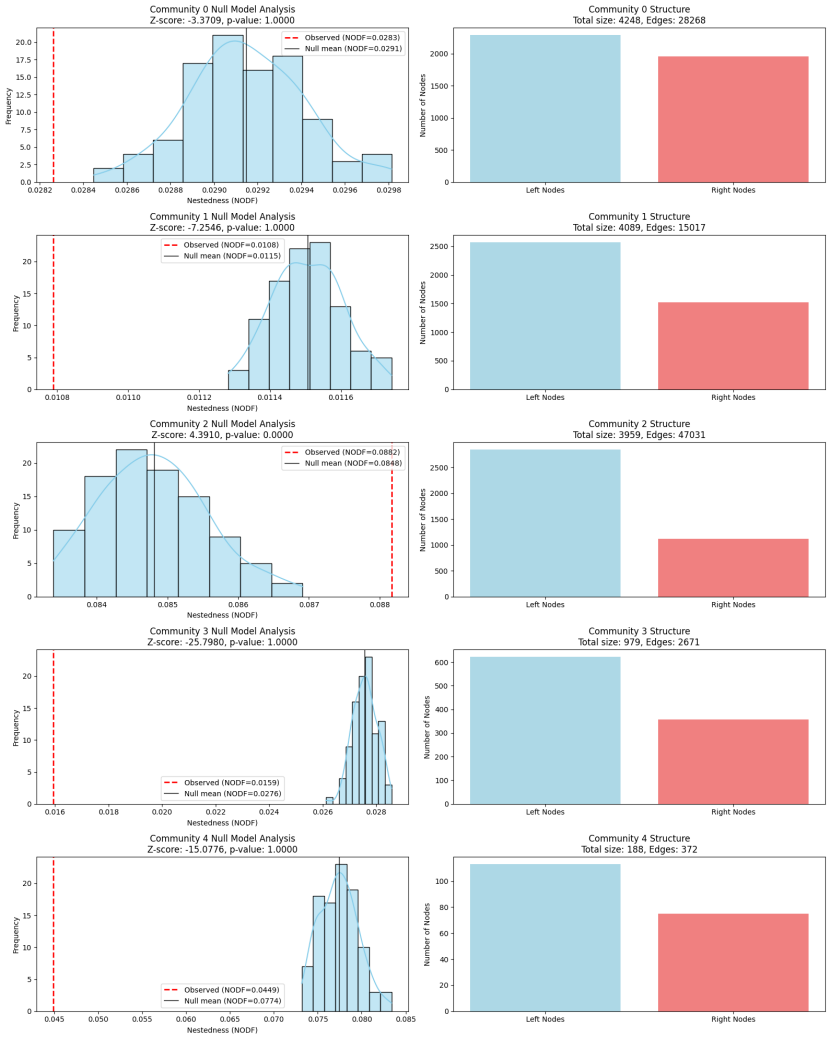


Figure 1: Comparison of observed versus null model nestedness scores for the five largest investor communities. The diagonal line represents equal observed and expected values, with points above the line indicating higher-than-random nestedness. "Left" stands for late stage investors, and "Right" for early stage ones.

This community displays a hierarchical investment structure wherein less-connected investors maintain co-investment relationships with a subset of partners associated with highly-connected investors.

2.4 Community Characterization

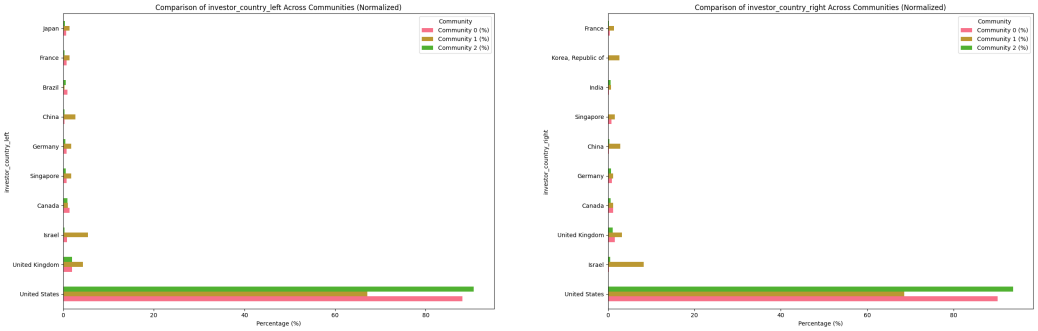
Analysis of the geographic and sectoral characteristics of the most nested communities reveals distinct patterns:

Mention only first 3 communities are being compared

2.4.1 Geographic Distribution

Geographic analysis of investor communities reveals distinct spatial clustering patterns that differentiate between early-stage and late-stage investment networks. Figure 2 demonstrates asymmetric geographic distributions across the bipartite network structure.

Better format geographic distribution figure



(a) Late-stage investors geographic distribution

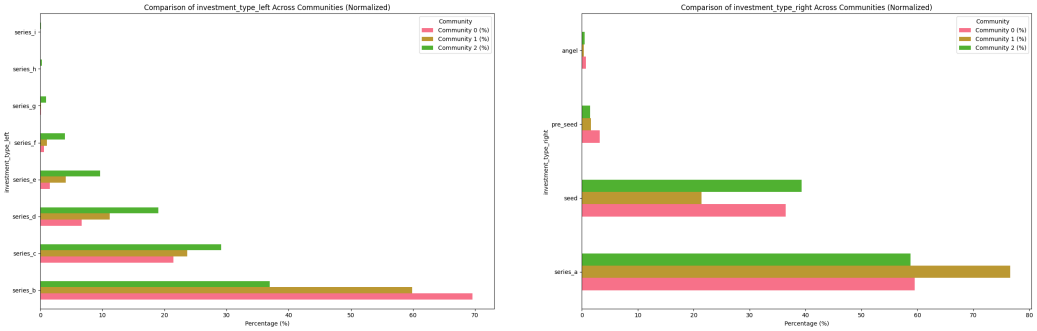
(b) Early-stage investors geographic distribution

Figure 2: Geographic distribution of venture capital investors across the largest communities. The bipartite structure reveals differential geographic clustering between late-stage (left) and early-stage (right) investor networks, with investors from Community 2 exhibiting greater international diversification.

Comment geographic distribution

2.4.2 Investment Stage Preferences

Investment stage analysis reveals systematic differences in funding round participation across communities. Figure 3 demonstrates the distribution patterns of investment types within the bipartite network structure, highlighting distinct preferences between early-stage and late-stage investor groups.



(a) Late-stage investment types distribution

(b) Early-stage investment types distribution

Figure 3: Investment stage distribution across the three largest communities. Late-stage investors (left) demonstrate concentration in Series B and later rounds, while early-stage investors (right) exhibit predominant participation in seed and Series A funding rounds. The distribution patterns reveal stage-specific specialization within investor communities.

Comment investment stages distribution

2.4.3 Sectoral Focus

Sectoral analysis reveals distinct industry specialization patterns within nested communities. Figure 4 illustrates the distribution of investment focus across technology sectors, demonstrating how different communities exhibit varying degrees of sectoral concentration.

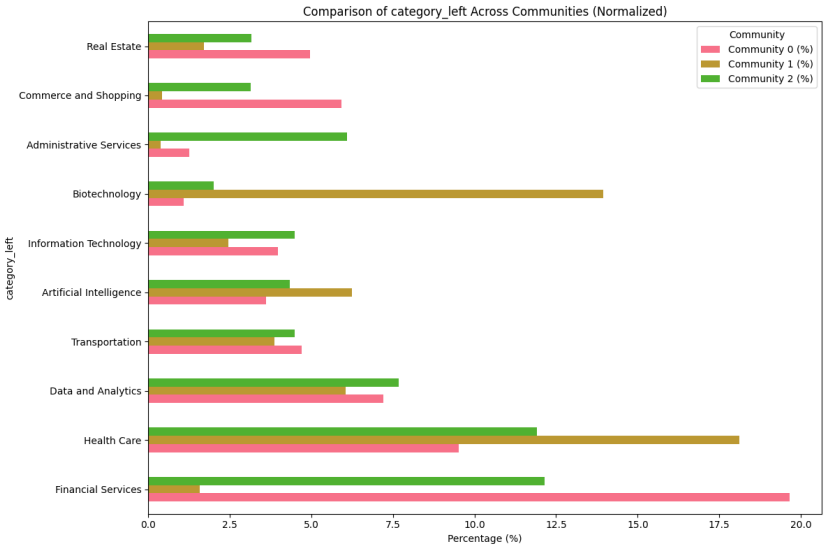


Figure 4: Sectoral distribution across the three largest investor communities. The analysis reveals differential industry focus patterns, with certain communities demonstrating concentrated investment strategies in specific technology sectors while others maintain broader sectoral diversification.

Comment sectorial distribution

2.4.4 Funding Characteristics

Funding analysis demonstrates differential investment patterns within nested communities compared to random network configurations. Figure 5 reveals systematic variations in funding amounts and investment frequency across community structures.

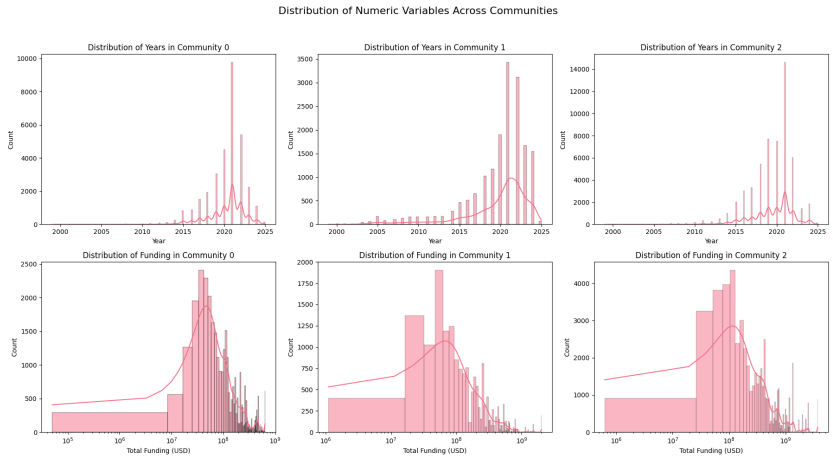


Figure 5: Funding characteristics across the three largest investor communities. The analysis reveals systematic differences in investment amounts, round frequency, and funding patterns between communities, with nested structures exhibiting distinct capital deployment strategies compared to randomly organized investor groups.

The funding characteristics analysis indicates that nested communities exhibit concentrated capital deployment patterns, with higher-degree investors participating in larger funding rounds while maintaining broader portfolio diversification. This pattern suggests efficient capital allocation mechanisms within hierarchically organized investor networks.

Comment more on funding characteristics

2.5 Implications and Future Directions

The discovery of significantly nested communities within the venture capital network has important implications for understanding investor behavior and startup access to capital. The hierarchical structure observed in these communities suggests the existence of informal investment hierarchies that may influence funding accessibility for entrepreneurs.

The presence of nested structures challenges the assumption of random mixing in venture capital markets and suggests that certain investors may serve as "gatekeepers" who influence access to broader investment networks. This finding aligns with social network theories about structural holes and brokerage positions [2].

The identification of these nested communities opens several avenues for future research into the social and economic mechanisms that drive venture capital ecosystem organization. Understanding these patterns may inform policy discussions about startup ecosystem development and investor network formation.

This analysis provides the foundation for deeper investigation into how nested investor communities influence entrepreneurial ecosystems and capital allocation efficiency, which will be the focus of subsequent research phases.

Investigate the economic consequences of nested community structure on startup success rates and funding efficiency.

Analyze the temporal evolution of community nestedness to understand how these structures emerge and persist over time.

Examine whether nested communities provide better or worse outcomes for portfolio companies compared to random investment patterns.

Apply social network theories of structural holes to understand the role of highly connected investors in nested communities.

Investigate whether the nested structure reflects information asymmetries or risk-sharing mechanisms among investors.

Develop theoretical models to explain the emergence of nested structures in investment networks.

Compare nestedness patterns across different geographic markets and time periods to understand generalizability.

References

[1] Mário Almeida-Neto, Paulo Guimarães, Paulo R. Guimarães Jr, Rafael D. Loyola, and Werner Ulrich. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, 117(8):1227–1239, 2008.

[2] Stephen P. Borgatti and Daniel S. Halgin. On network theory. *Organization Science*, 22(5):1168–1181, 2011.

[3] Jean-Michel Dalle et al. Accelerator-mediated access to investors among early-stage start-ups. *Research Policy*, 2025. In press.

[4] Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesús San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5(1):4114, 2014.