The Business Social Networks of Indian Unicorn Startups

MS in Applied Information and Data Science, HSLU

Course: Analysis and Modelling of Social Interactions

Introduction & Problem Definition

The market system of Capital investment, embodied by Venture Capital (VC) firms, collectively thrives on the principle "The Rich Get Richer", known as the Matthew Effect [ES01] :

> For to everyone who has will more be given, and he will have an abundance. But from the one who has not, even what he has will be taken away.
>
> *Matthew 25:29 (ESVUK)*

In Network Theory, this effect could be explained by Preferential Attachment - the tendency of higher-connected nodes to accumulate disproportionately more connections based on how many connections they already have. It is a given that VC firms with more cash on hand can make larger investments with larger absolute returns - their returns scale with the size of their bank balance.

That is true for their Financial Capital, but does the same principle also apply to their Social Capital? Specifically:

- How do companies and investors' connections within the business social network scale?
- What are the differences between the combined social networks of Companies-Investors vs. Company-Company or Investor-Investor networks?
- Which ones show evidence of Preferential Attachment?
- Do their networks look more like random networks, or more like Scale-Free networks — ones following a 'Praeto Distribution' or Power-Law with a hub-and-spokes centralized pattern?
- Do these networks between business entities resemble real-world human networks, such as by displaying Small-World [WS98] properties (comparably short path lengths but higher clustering than random networks of the same size)?
- What generalisations and long term strategies could companies and investors adopt, using insights from Network Analysis of their business social network?

Investors invest in groups ('Syndicates') alongside other people they know and trust (rewards associative similarity). But they also try to scale up the size of their investments and profits, and gain access to new exclusive information about new markets and opportunities (rewards variability)

Economies are complex systems, so the development of a business social network is driven both by random chance factors, and intrinsic self-organising predictable factors. By analysing an Indian VC dataset of recent Unicorn startups (companies which have surpassed $1bn in valuation [FS23]), we aim to provide insight into these questions, by identifying certain intrinsic factors emerging from the connectivity of this social network, while referencing comparable simulated random network models as control benchmarks.

Problem Definition

This investigation can be formalised by seeking clarification of the following hypotheses:

- **Hypothesis 1:** The company network is unlikely to develop hubs so will not demonstrate scale-free or preferential attachment characteristics (due to its more strictly bounded degree distribution).
- **Hypothesis 2:** The investor network will develop hubs and thus exhibit scale-free and preferential attachment properties (due to its less bounded degree distribution).
- **Hypothesis 3:** All business networks will display small-world properties (due to being super-structures on human social networks).

These hypotheses were selected to constrain the scope of the study within practical limits. Additional research avenues (i.e. community detection, link prediction, feature importance) are discussed in 'Further Work'.

Background Research

Motivation for Social Network Analysis

Network analysis provides quantitative approaches to investigate answer pressing questions about social networks. Research applying systems control theory to the neural network map of the Caenorhabditis Elegans worm [YV17] has shown it is possible to identify network structure-function relations and find the right nodes to control to steer overall dynamics. Network analysis can also reveal that despite women holding only about 9-13% of all board seats globally, compared to men (in a dataset of inter-linked company board membership) they have higher average degree and higher betweenness centrality [EG20], making them on average better networked.

Counter-intuitively, despite being fewer in absolute number, their more central connectivity in the overall network is empirical evidence that women's participation in these networks is nontheless substantial and influential, challenging potential assumptions about gender stereotypes in corporate governance. This not only validates the use of social network analysis in this context but also emphasizes its critical role in uncovering and understanding complex, often hidden, network dynamics.

Venture Capital co-invest in groups (syndicates) because this reduces individual investment risks and expands access to diverse expertise and larger networks. But it also brings increased coordination costs and potential misalignment among partners [LN23]. These business social networks facilitate deal sourcing and information sharing, but can also create biases based on prior affiliations [SC23], such as favortism and preferential contact among alumni graduates from the same university [GM23]. This highlights that beyond correlational studies, in causal analysis (modelling 'What really causes connections and business decisions to be made?') human and social factors will most likely also influence investment decisions, not just financial fundamentals or merit.

Models in Social Network Analysis

Social Networks can be characterised according to what conditions their structures satisfy (all examples and definitions from this section are reproduced from [BA16] "Network Science" Chatper 4: "The Scale-Free Property")

Random Networks: In network theory, simulating random networks serves as a baseline to understand more complex structures. Purely random networks can be 'grown' iteratively by nodes randomly (e.g. the Erdős–Rényi model), or semi-randomly while still obeying specified constraints like with Exponential Random Graph Models (ERGM) [RP07] [GJ16] to mimic certain properties of the original network under study. The resulting network gives a baseline providing a null hypothesis for statistical comparison. Such simulations help in understanding the likelihood of certain network characteristics emerging by chance. When we evaluate real-world networks against this baseline, we can better appreciate the significance of observed structures, to identify which properties deviate from what would be expected to be observed purely by random chance. The expected occurrence of high-degree hubs in fully random networks is vanishingly small.

Scale-Free Networks: These are networks where the nodes are not connected randomly, instead the degree distribution follows a power law meaning there are no typical nodes, but rather a continuum from low to high connectivity. Nodes are connected in such a way that most nodes have few connections, but a handful of nodes (called hubs) can have many connections. This effect is best observed with larger datasets where the negative exponent of the power-law for a typical scale-free network is typically between 2–3 [BA16]. The network is robust against random failures; if a non-hub node fails, it doesn`t affect the overall connectivity of the network significantly. However, if a hub is targeted and fails, it can cause substantial disruption. Notably (from [BA16]) "For the scale-free property to emerge the nodes need to have the capacity to link to an arbitrary number of other nodes. ... The scale-free property is absent in systems that limit the number of links a node can have, effectively restricting the maximum size of the hubs."

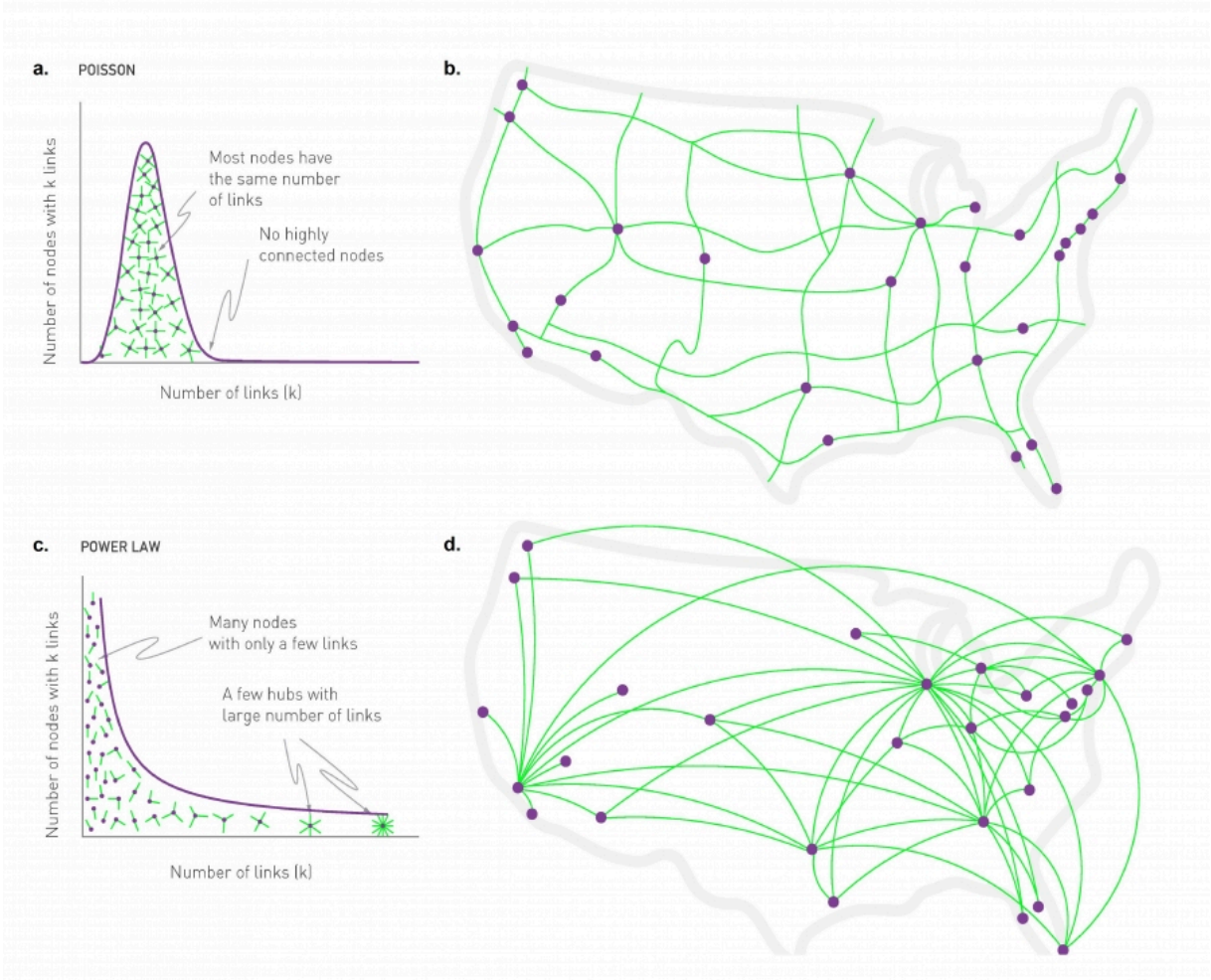


Illustration from [BA16] illustrating the connectivity of a simple random network (top) vs. a power-law distributed network (bottom), illustrating how a few cities can become airport hubs because the maximum number of connections by air is greater than by highway.

Preferential Attachment: This is one formation mechanism which can lead to the growth of networks with a scale-free structure. It characterises situations where new nodes are more likely to form connections with already well-connected nodes. This 'rich get richer' phenomenon means that the more connections a node has, the more it attracts, leading over time to the power-law distribution of node degrees that is characteristic of scale-free networks. This is a scaling property, thus it requires datasets of sufficient scales to observe conclusively — in practice datasets of the order of $\sim 10^2$ nodes (such as those in this report) are somewhat too small to give definitive preferential attachment results [BA16].

Small World Networks: Many networks, including many real-world social networks, can be characterised as being small worlds (after [WS98]) characterized by a simultaneous combination of high clustering and short average path lengths, which means that while nodes tend to create tight-knit groups with high interconnectivity, any node can be reached from any other through a small number of steps. This gives an efficient structure for information or resource transfer, indicating that the social world is much more connected than one might intuitively expect, because groups are tight-knit, but also have enough randomness interconnecting them to spread messages quickly.

Research Methods

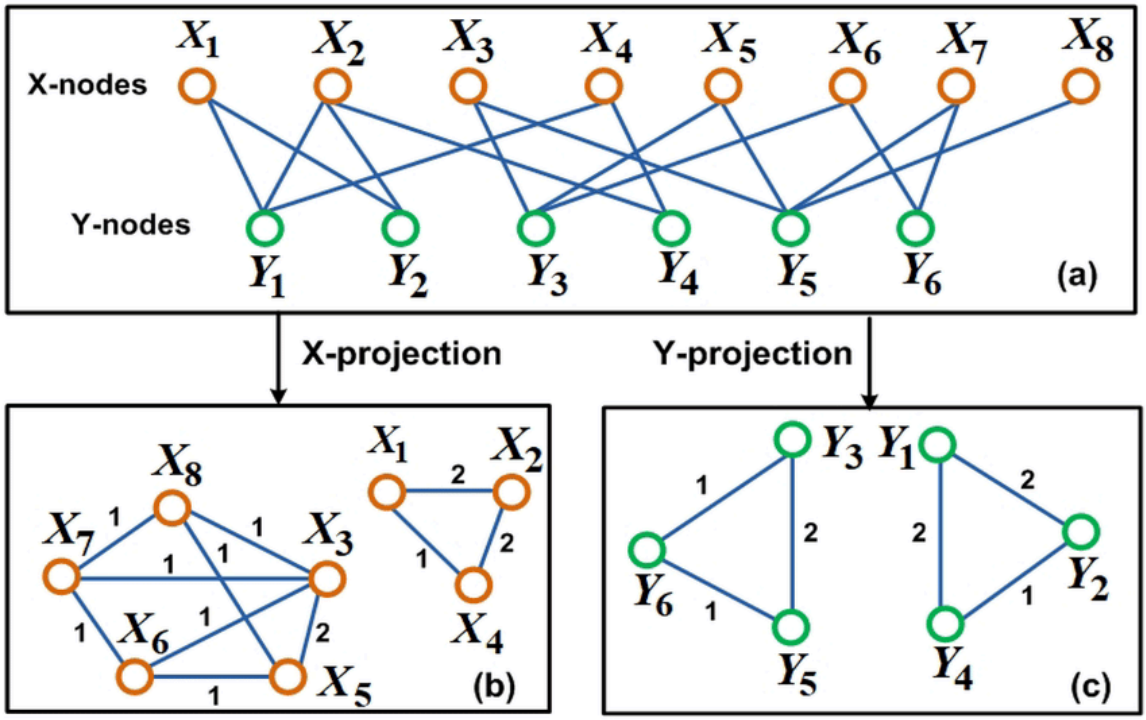
Data Sourcing & Enrichment

This study examines a dataset from Kaggle sourced from this dataset consisting of selected investors in Indian Unicorn start-ups. This contains approximately 100 companies and 150 investors. Which investors got chosen per company is an opaque process and was manually capped at 7, though there may be more. Likewise which companies a given investor has invested in are only limited to this Unicorn sample, so this is a very narrow sample and not representative of the overall market, or of all of the investments taken by a given investor.

Subscriptions to Venture Captial data brokerages are cost-prohibitive, so to gain more context on the investors' founding year and location, the OpenAI and APIFY APIs were used to scrape the top 20 results for a given investor on CrunchBase.com, automatically select the most relevant one, and extract the relevant details (the GoogleMaps API was also used to extract location information). Companies' industry categories were binned into a smaller number of more comprehensive categories, as were Investors' locations. The full final list of enriched features is visible in Appendix I: Final Enriched Features List, many of which are employed for EDA and tooltip annotations, but the only ones used for the consequent network analysis was the list of company-investor links.

Calculations

The graph connecting Investors to Companies is modeled as a (unidirectional) bipartite graph. Many networks can be modeled as instances of the bipartite graph category [NS01], including: academic co-author publications, musical artist collaborations on songs, house / yacht group rentals, actors co-starring in movies, coaches relating to players via teams, directors being on multiple boards of directors etc... By treating this investigation at the level of abstraction of analysing a bipartite graph, many techniques can be employed from fields outside of VC, and explanations and insights from this analysis can in turn be extrapolated back to broader fields.



An example of remapping a combined bipartite network graph down to two univariate sub-networks [GM22] e.g. in the X-Y network, nodes X_1 and X_2 share no direct X-X links, but are both indirectly connected via node Y_1 and Y_2 , which becomes a link with weight 2 in their unipartite X-X network.

The asymmetric nature of the company-investor relationship reflects broader market bipartite-network patterns like buyer-seller behaviors (in Venture Capital, investors are buyers and companies are sellers of investment offers). The overall bipartite company-investor network was projected down to its unipartite company-company and investor-investor components. These 3 networks also had additional binning applied, to collect all investors that only appear once as One-Time investors (either as a single entity, or sub-categorized by their investor region).

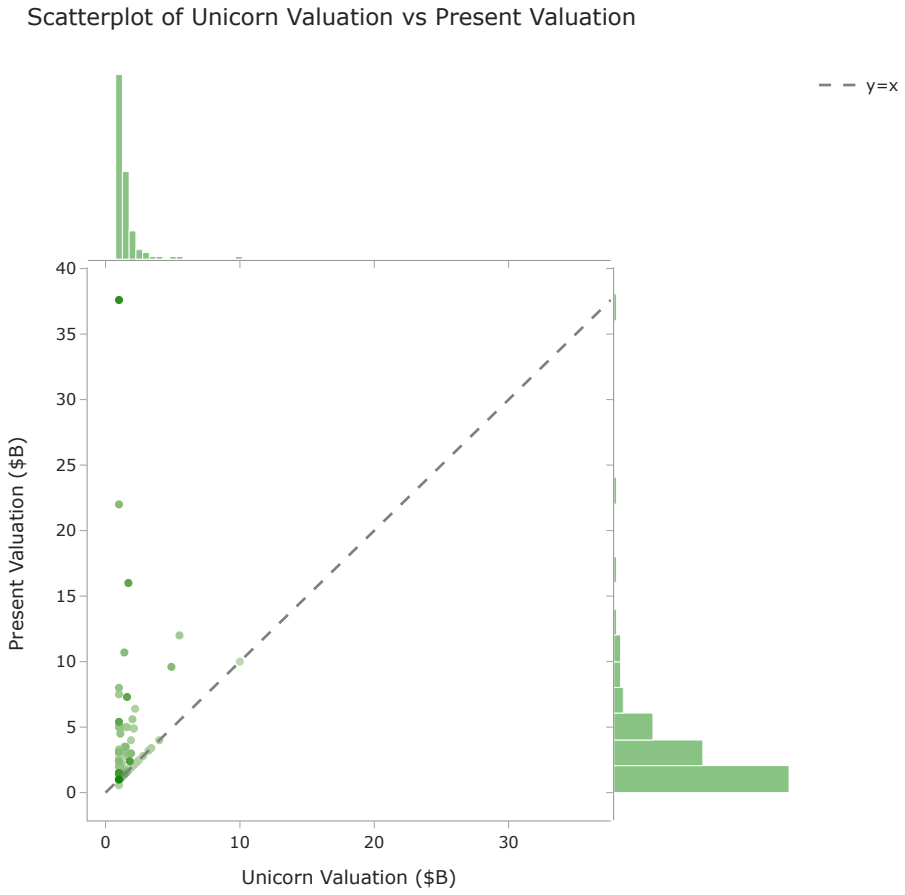
The Networkx package was used to calculate the standard network measures for the largest sub-component of the graph (which immediately becomes giant component fully connected), for each iteration of the graph as nodes get added in their timeline sequence of becoming Unicorns: degree distribution, degree centrality, betweenness centrality, clustering coefficients, closeness centrality, pagerank centrality, mean shortest path length, log-log degree distribution with linear fit (done according to [BA16]), and preferential attachment score (calculated using Networkx according to [LK04]).

The Rk value was also calculated (in batched timestep stages of network growth to not be too sparse and granular) according to [NE01], which is an alternate measure of preferential attachment, comparing how the probability of new edges attaching to high-degree nodes grows with time. If this graph shows a positive trend, it indicates that as nodes get more connections (higher degree) they have a bias towards connecting to high-degree nodes, indicating a preference for high degree nodes i.e. preferential attachment.

Lastly, the 3 main types of network were characterised by calculating their scale-free (log-log) slope, Rk (preferential attachment) score, clustering coefficient, and average shortest path length. For each network, the same metrics were calculated and averaged for 100 simple random graphs for comparable reference baselines, simulated using approximately the same number of nodes and edges as their original target graphs.

Exploratory Data Analysis

Company Valuations

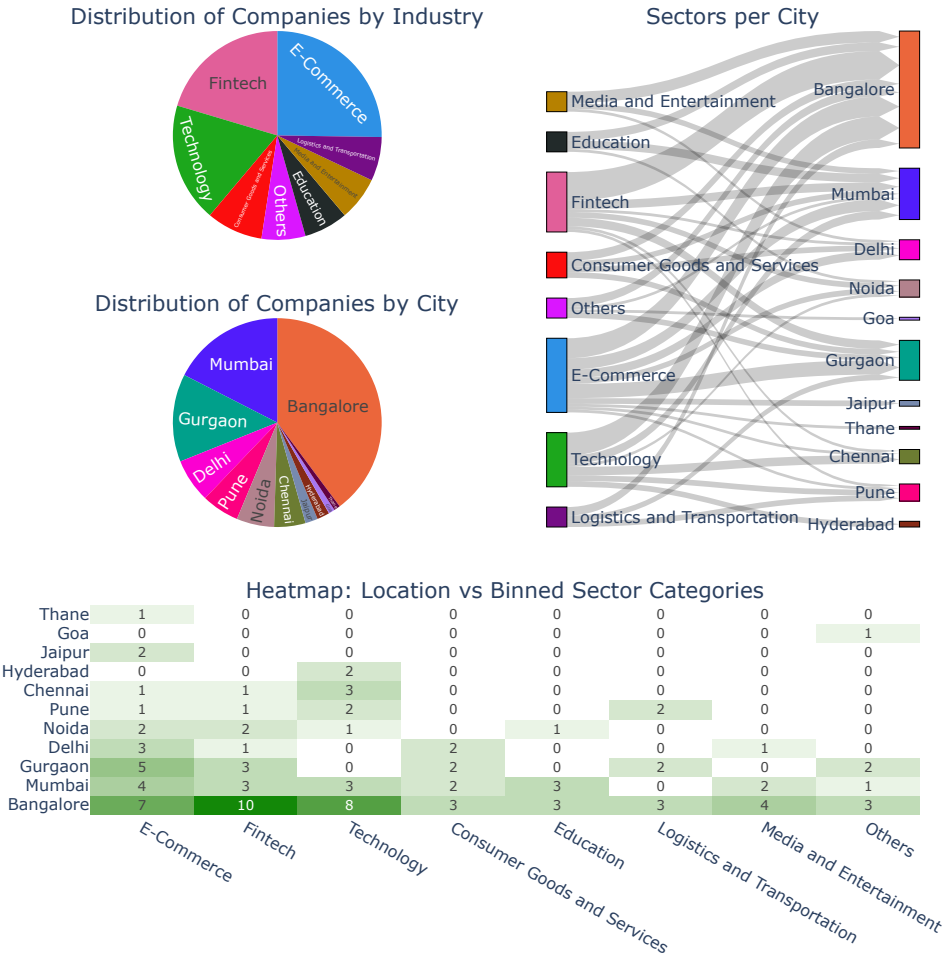


Comparison of company valuation at the date they passed the $1bn 'Unicorn' mark vs. present valuation. Darker plots are older (hover or zoom to view details).

Despite a few outliers, most Unicorns in this dataset are recent thus do not (yet) show long-term growth. Absent one outlier, all Unicorns in this time frame maintain the same value as their initial entry, or have grown in value. These results are primarily skewed by recency bias and survivorship bias, before they have had time to fully mature or fail.

Detailed data is not available to break down which investors invested at which stage, for which evaluation value. This information is typically curated by Investment data brokers including Pitchbook.com and Crunchbase.com , and require very high annual subscription fees to access. Thus, although this financial information would be one of the most important signals for Venture Capitalist investors (who are in the business of making a profit) it will not be analysed further in this report.

Company Sectors and Locations

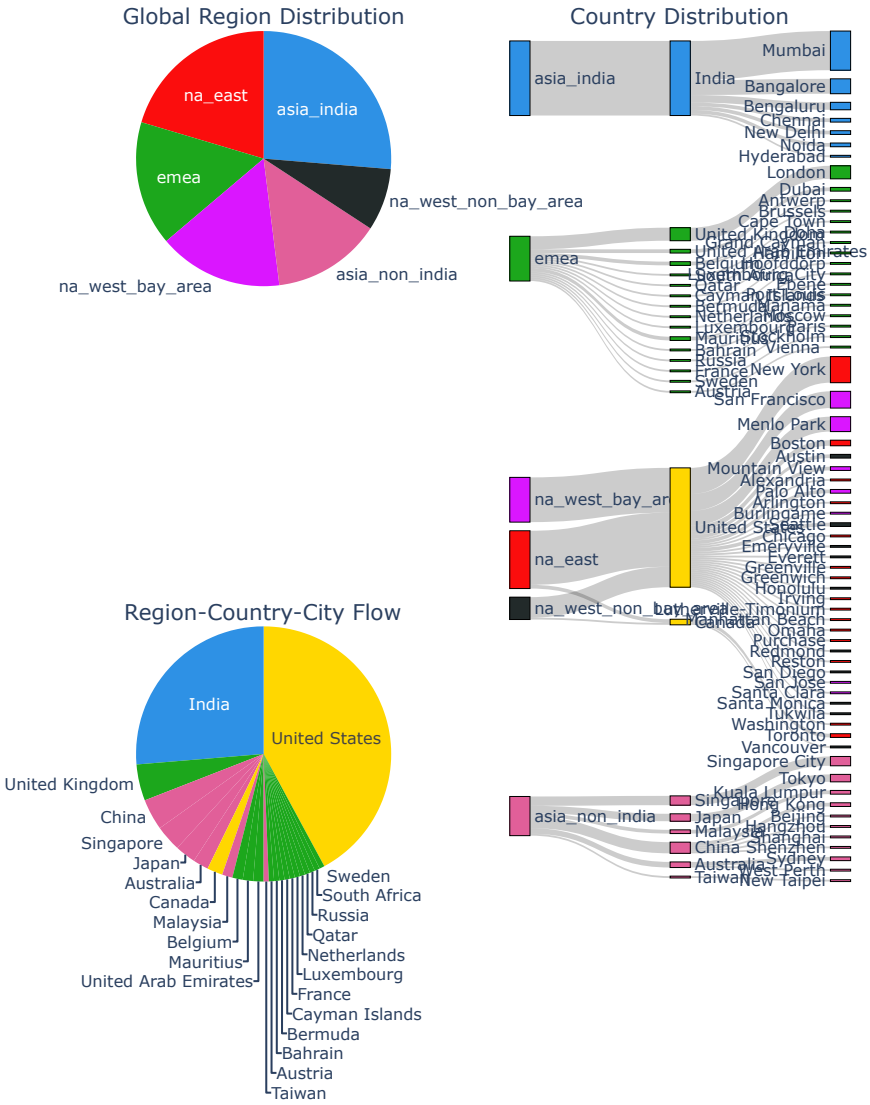


Charts showing how Companies' Sectors (specially binned) and Locations interact (hover for detailed information).

These sectors were carefully binned for a simpler overview. It bears reminding that this dataset only filters for the unicorns, the unusually high-performing companies. Thus it should only be taken to reflect truths about the subset of Indian startups in these various sectors and locations who happened to become very successful. This is by no means a representative sample of all startups in India.

E-Commerce, Fintech, and Technology are the most prominent sectors, mostly occurring in Bangalore, indicating a concentration of expertise in that location. Bangalore's large slice of the city pie chart combined with its prominence in the Sankey diagram suggest it is a critical center for startups (for example home to PhonePe, known for its user-friendly interface and a massive user base of 400 million), which reflects the reality of the Indian market. It also hosts a greater diversity of sectors compared to the runner-ups Mumbai and Gurgaon. This suggests successful companies are concentrated in sector-specific regions, and investors may adapt their targeting strategies accordingly.

Investor Locations

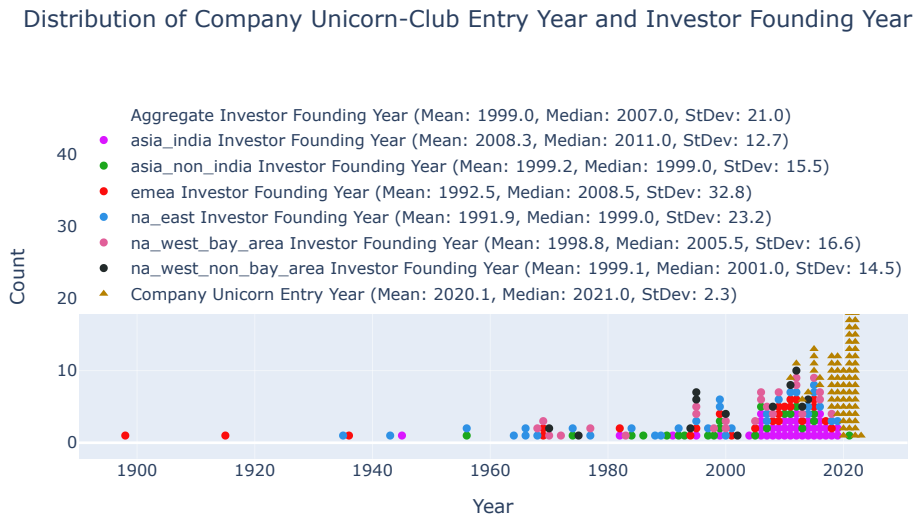


Charts showing how Investors are organised by Global Region (specially binned), Country, and city (colors match per region, hover for detailed information).

There is a high concentration of Venture Capital investors originating in the US (especially New York, San Francisco, and Silicon Valley i.e. Menlo Park & Mountain View), which reflects global market trends. The manually selected regional bins yield an even distribution of investors across the various regions.

This study focusing on India explains the large proportion from India. Mumbai takes a greater slice of the (Indian) investors (~50%) compared to its ratio of unicorn companies (~20%). This indicates Mumbai is a stronger hub for investors, contrasting with Bangalore as a hub for companies.

Year Investor Founded vs. Year Company Became Unicorn

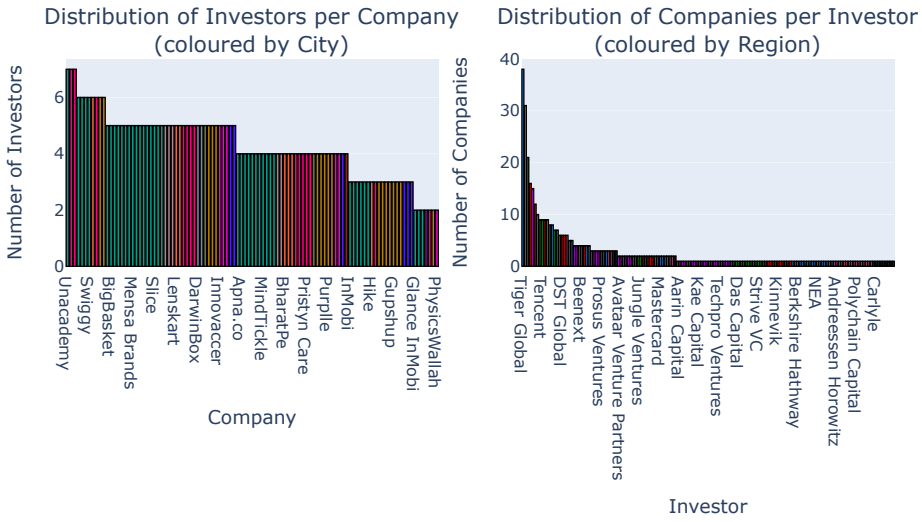


Each brown triangle represents a company, the rest represent investors (hover for more info, click and drag to zoom).

Most investors are established before unicorn companies emerged. The investors from the East Coast of North America are the oldest (average 1992). The Indian investors are the most recent (average 2008) suggesting that the investment ecosystem in India is relatively younger. The EMEA investors have the greatest spread of ages (stdev = 33yr).

Effectively all of the unicorn companies are founded after investors, allowing us to only use the Unicorn year of companies to summarize the entire time sequence of investment. The data may imply that investors are at an advantage from being long-established, giving more time to strengthen connections, experience, and capital reserves, relative to unicorn companies.

Number of Investors per Company & Number of Companies per Investor



Node Degree i.e. how many connections each Company, and each Investor has, color corresponds to city and binned region respectively (hover and zoom for individual detail).

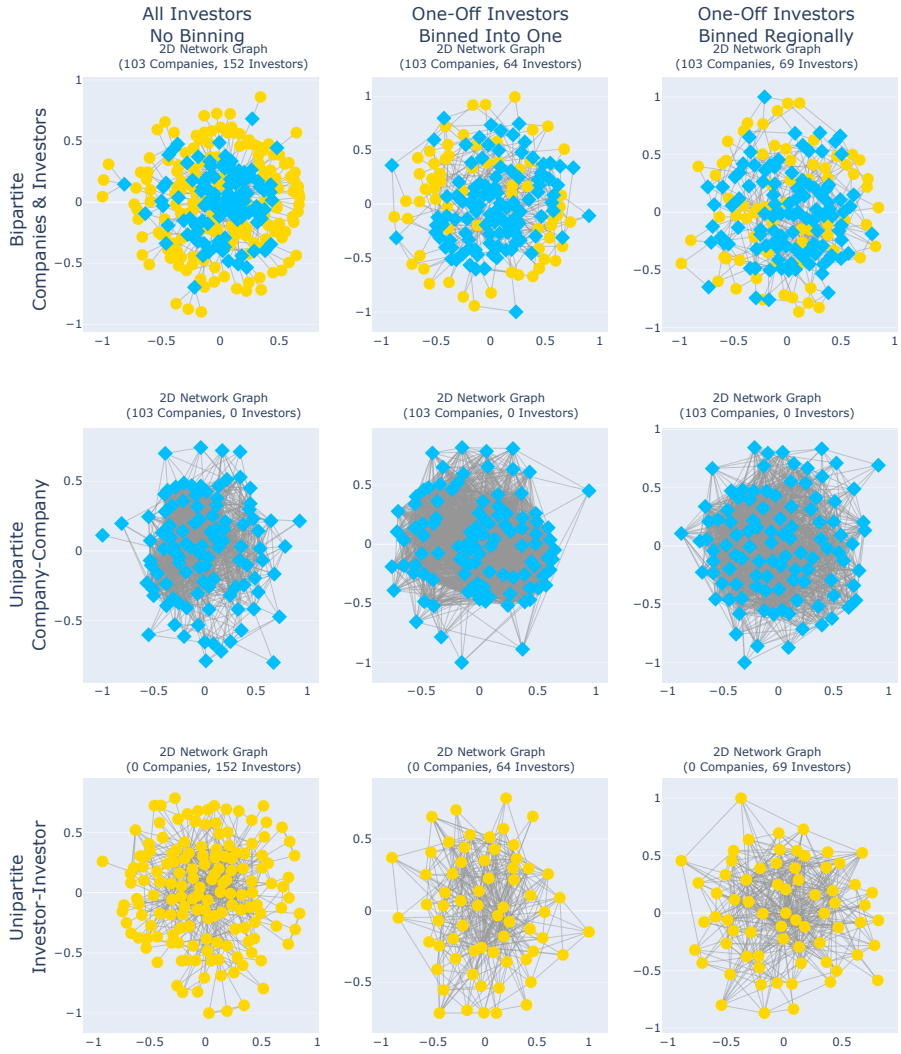
Most companies have a similar number of investors, indicating a balanced investment landscape where companies are able to attract a consistent level of interest from investors, evenly across cities.

In contrast there is significant variation in the number of companies per investor, with a heavily right-skewed distribution. The viewport can be zoomed by dragging horizontally to show exactly which investors these are. Conversely a majority of investors only invest in few companies, or even just one company i.e. are One-Off investors. To consolidate their contributions these are sometimes binned by region in subsequent analysis.

This difference between the degree distribution of the Investor vs. the Company nodes reflects a fundamental difference between their dynamics — there is an upper cap on how many investors a company can realistically receive (in this case 7, though this may be somewhat reduced due to the sampling methodology), due to diminishing benefits from adding extra investors. In contrast successful large-scale investors can continue to invest in many new companies each year, which compounds the scale of their portfolio and number of investments they can tend to.

Network Graph Results Visualization & Analysis

As shown from the degree distributions and other exploratory data analysis, the way **companies** are connected in the network is fundamentally different from the way that **investors** are. This is best seen by exploring the dashboards below, where the individual nodes' values are shown (zoom & hover) along with the context of those values' overall distributions.



Network | Degree Dist. | Log-Log Degree Dist. | Degree Cent. | Betweenness Cent. | Clustering Coeff. | Closeness Cent. | PageRank Cent. | Shortest Path

Detailed grid of all networks's values for a given measure, selectable by clicking the buttons (additional information is available on zoom and hover, double-click to reset axes).
EDIT: Replaced interactive graphc with static image for performance reasons.

Overview grid of all measures for a given network, selectable by dragging the slider (additional information is available on zoom and hover, double-click to reset axes).
EDIT: Replaced interactive graphc with static image for performance reasons.

The overall Bipartite network has been split into its Unipartite Company-Compnay and Investor-Investor projections. These have also had their One-Off investors binned, either into One node, or into Regional nodes. This is to clean up the data and prioritise the high-degree investors, while keeping much of the network connectivity. These One-Off variants show less skewed clustering coefficients (due to more even degree distribution), and shorter mean shortest path-lengths (because there are fewer nodes overall so fewer paths to take).

Network: Binning One-Off investors has a significant impact on the Company-Investor and Investor-Investor networks, but not on the Company-Company network, reducing visual clutter. Zooming makes clear there are certain investors with very high degree, and many other high measures as a secondary result.

Degree Distribution: As previously shown, the number of connections to companies is distributed quite differently to the connections to investors. Investors, and the overall bipartite network, have heavily sekewed degree distributions (median < mean), whereas companies much more evenly spread distributions (median ≈ mean) with much larger standard deviations. This means investors are more likely to display scale-free and preferential attachment formation dynamics, wheras the Company network will be more uniform and more like a random network.

Log-Log Degree Distribution: This graph will be used to evaluate scale-free properties, which are evident for networks involving investors. Companies' fits do not fit a negative power law (evidenced by their small positive slope and small positive r value), but ones involving investors do (moderate negative slope, large negative r value). While these slopes are not within the expected power range of 2–3 for a typical scale-free network [BA16], they are nonetheless statistically significant and tend towards indicating scale-free properties for the Investors networks.

Degree Centrality: As a consequence of the skewed Degree distribution, the Degree Centrality is strongly skewed for the Company-Investor and Investor-Investor networks. Companies have much higher mean Degree Centrality than investors, indicating they are overall more highly interconnected.

Betweenness Centrality: The Betweenness Centrality is on average low, with a few outliers corresponding to highly connected nodes. The time-evolution of this measure is shown below, showing that the outliers are only present for the Investor networks, and correspond to the high-degree nodes. These serve as bridges in the Investor networks, corresponding to high Betweenness. The Company-Company network is already so highly connected its Betweenness is quite low, no individual nodes emerge as being bridging or path-critical, with an absence of structural holes.

Clustering Coefficients: Clustering is calculated slightly differently for the Bipartite graph (because by definition it will have no triangles) and is on average lower than for the Unipartite networks, which have quite high levels of clustering (relative to a random graph baseline) which provides strong evidence for Small World structure, as will be discussed below.

Closeness Centrality: The Company-Company network has higher average Closeness than networks containing investors, and the closenesses are increased upon binning One-Off investors, which effectively shrinks the network diameter.

PageRank Centrality: By zooming it is evident there are a few outlier investor nodes with very high page-rank, which again, correspond to the nodes with highest degree and betweenness — the most influential bridging nodes.

Mean Shortest Path Length: Because there are far fewer nodes in the Unipartite networks, those networks' absolute values and overall network diameters are smaller than for the Bipartite network. Binning One-Off investors reduces the number of leaf nodes, which shortens the overall max path length (10 down to 7), which removes hard-to-reach peripheral nodes i.e. the companies or investors which will have the hardest time finding mutual contacts.

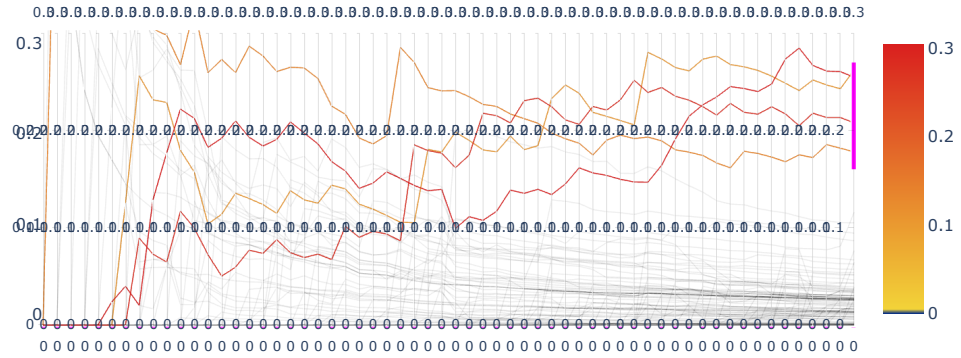
Evolution of Measures Through Time

These network measures were also calculated over time, as new nodes were added to the network, generally remaining quite steady over time. The two measures which showed diverging or non-monotonic trends are shown below. Click either 'Betweenness Centrality' or 'Preferential Attachment' labels to toggle which sets of timelines to display.

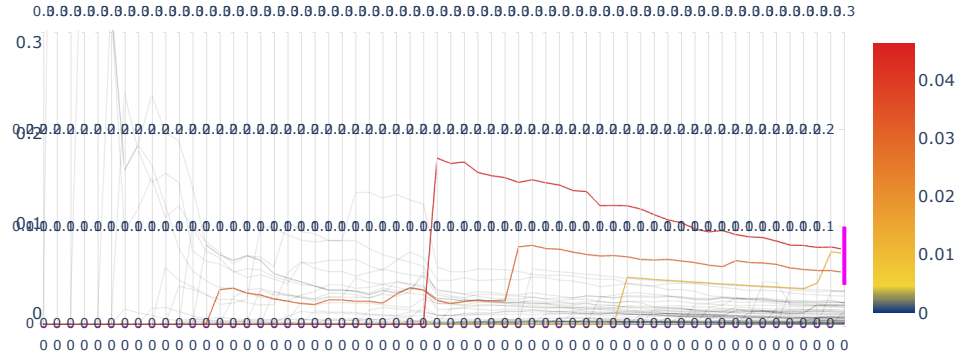
Betweenness Centrality

Preferential Attachment

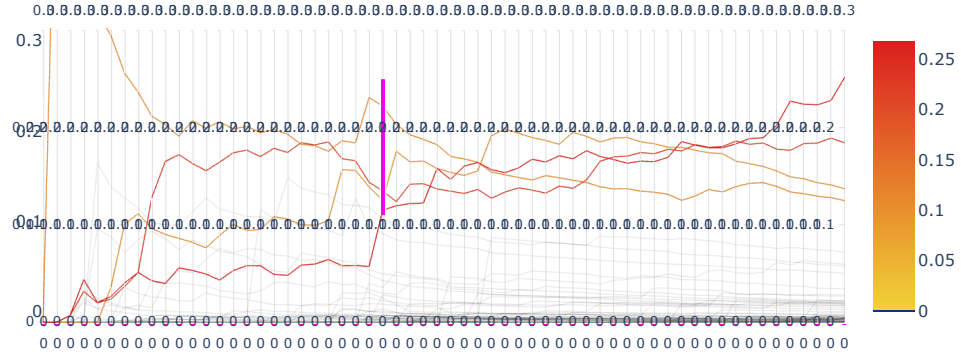
Bipartite Company-Investor
Betweenness Centrality is spread very unevenly



Unipartite Company-Company
Betweenness Centrality is spread somewhat unevenly



Unipartite Investor-Investor
Betweenness Centrality is spread very unevenly



Evolution timelines of Betweenness Centrality and Preferential Attachment measures, showing pre-selected filter ranges in purple (drag bars or click & drag lines to adjust filters) - color-scales are centered on final outcome median values, vertical scales are equal per measure.

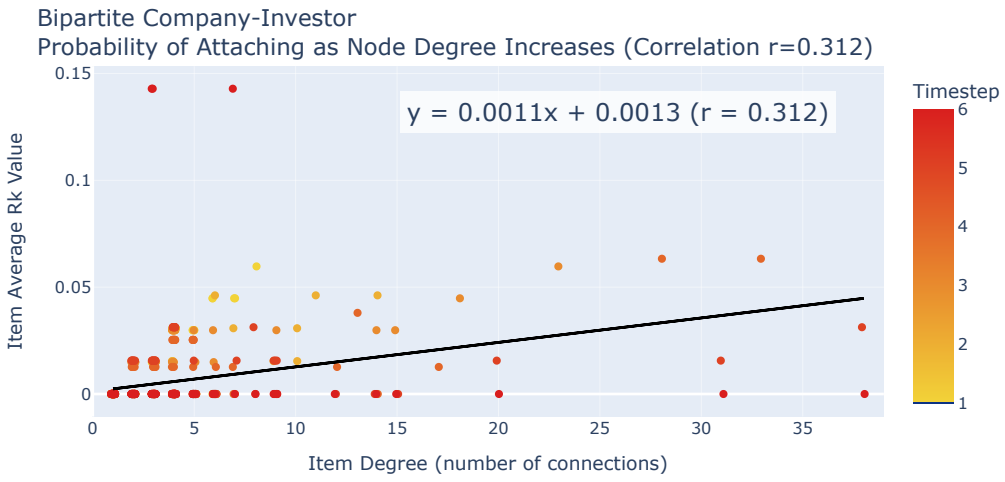
Betweenness Centrality: The color scales indicate the final outcomes are extremely skewed, with only a few upper outliers. For the overall network and the Investor-Investor network the nodes with the highest final betweenness all start gaining attachments from the very beginning. For the Investor-Investor network this this is a strong effect, where the top outliers soon emerge and stay in front even from halfway through the series (drag purple filter slider vertically to explore). In contrast the Company-Company network is less skewed, and its bridging outliers do not exclusively emerge from the first few timesteps. Overall this indicates a few Investor nodes emerge early as the bridging nodes with consistently high Betweenness Centrality.

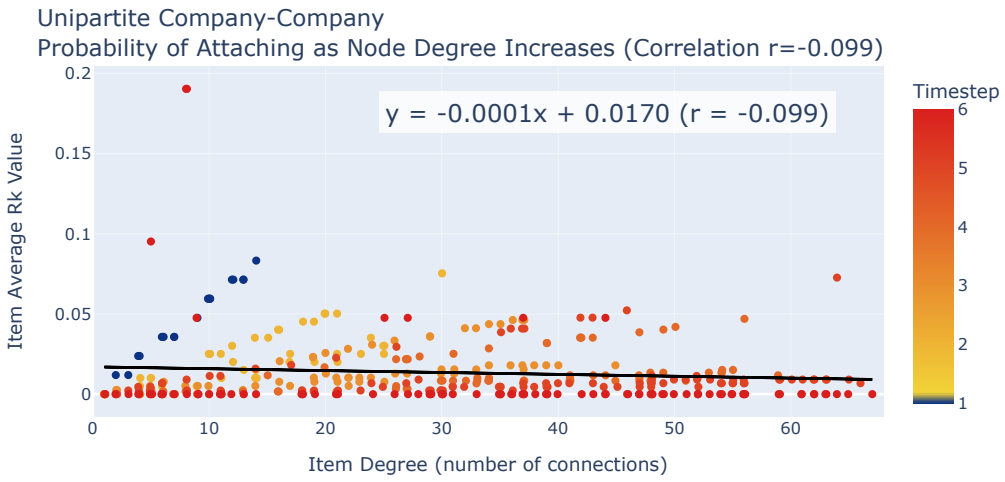
Preferential Attachment: The Preferential Attachment score calculated over time shows the Company-Company network having an even distribution of scores, and highest final scores from nodes that join the network across the whole timeline. The fact the absolute scores are higher can be explained by the way preferential attachment is calculated, giving higher scores for more connected nodes - leading to score inflation because most Companies are highly connected with other Companies (on average 34.5 ± 19.5 connections, vs. the Investor-Investor average of 8.1 ± 9.9 connections). Companies have In contrast the networks with Investors only show a few outliers, which again emerge quite early, they persist across time, and grow monotonically.

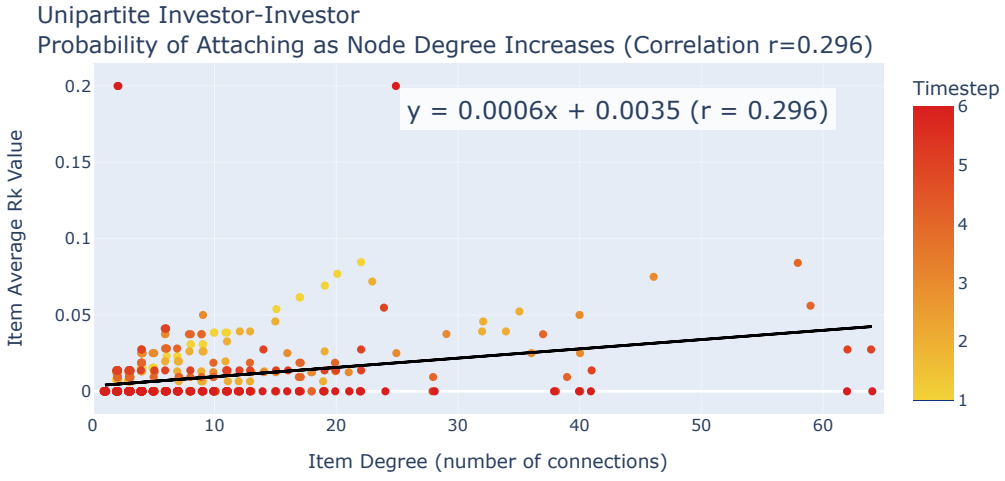
Overall, certain investor nodes attract more connections over time, reinforcing their centrality in the network, with a few nodes becoming increasingly dominant. This suggests those networks involve Preferential Attachment mechanisms. In particular, these also happen to all be the large institutional investors with high degree (Sequioa Capital, Tiger Global, Accel India, Steadview Capital). In contrast, the Company-Company network has high scores for companies joining at various times, suggesting a more uniform distribution of connections thus not preferential attachment.

Preferential Attachment and Rk Analysis

Trends in Rk values show the tendency of nodes to connect based on the degree of existing nodes, as shown below:







Correlation between how the degree of a node increases vs. that node's probability to attach to other nodes with high degree. Rk is calculated in batches denoted by color (hover for details).

The slope lines have only have absolute size due to the units and scaling used, thus this does not detract from their usefulness. The fit line for the Company-Company has a small negative slope and r value (-0.0001 and -0.099) indicating a deceasing trend, or un-preferential attachment, where if a node has higher degree, it is less likely to attach to other nodes with higher degree. However overall the correlation is not statistically significant, so only shows negligible patterns. Therefore this does not suggest that companies connect based on existing network degrees.

In contrast Investor networks show a positive moderately significant trend i.e. as nodes age and gain more connections in the network, they are more likely to attach to better-connected nodes than less-connected nodes. This is direct evidence of preferential attachment developing across time, but only for Investor networks.

Network Analysis: Scale-Free, Preferential Attachment, and Small-World

This section summarises the characteristics of our networks compared to reference baseline Random networks (averages of 100 simulation runs), focusing network properties (and their corresponding indicators): Scale-Free (Log-Log slope), Preferential Attachment (Rk slope), and Small-World (Clustering & Shortest Path).

Property	Bipartite	Bipartite Random	Company-Company	Company-Company Random	Investor-Investor	Investor-Investor Random
Number of nodes	255.0	255.0 (±0.0)	103.0	103.0 (±0.0)	152.0	152.0 (±0.0)
Number of edges	440.0	435.0 (±21.5)	1780.0	1780.0 (±31.7)	619.0	621.0 (±21.3)
Log-log slope	-1.55	-1.28 (±0.261)	0.0394	0.112 (±0.933)	-0.865	-0.0778 (±0.457)
Log-log R-value	-0.901	-0.685 (±0.0605)	0.074	0.0358 (±0.226)	-0.763	-0.0202 (±0.255)
Preferential attachment slope	0.00114	0.000938 (±0.000208)	-0.000117	-0.000688 (±5.64e-05)	0.000608	2.93e-05 (±0.000186)
Preferential attachment R-value	0.312	0.16 (±0.0321)	-0.0987	-0.337 (±0.0247)	0.296	0.00606 (±0.0394)
Clustering coefficient	0.231	0.0109 (±0.00511)	0.695	0.338 (±0.00634)	0.752	0.0536 (±0.00739)
Shortest path length	4.38	3.8 (±0.899)	1.73	1.66 (±0.00605)	2.51	2.61 (±0.0345)

Results comparing various measures to their equivalent baselines calculated from random graphs using roughly the same number of nodes and edges and timestep evolution sequences.

Network Type / Property	Bipartite Investors-Companies Network	Unipartite Companies Network	Unipartite Investors Network
Scale-Free	**Strong scale-free structure** Negative slope with strong r value (-1.55 with r=-0.901) steeper than random (-1.28 ±0.261 with r=-0.695 ±0.061).	**No scale-free structure** Slope is flatter than random baseline (0.0394 <) 0.112 ±0.933 both with negligible r values.	**Moderate scale-free structure** Negative Slope -0.865 (r=-0.763) steeper than random baseline (-0.0778 ±0.457 with negligible r value).
Preferential Attachment	**Weak preferential attachment** Slope slightly greater than random baseline (0.00114 with r=0.312 > 0.000938 ±0.000208 with r=0.16 ±0.0321).	**Lack of preferential attachment** Negligible slope indistinguishable from random baseline (-0.000117 with r=-0.0987 ≈ -0.000688 ±0.0000564 with r=-0.337 ±0.0247).	**Weak preferential attachment** Slope slightly greater than random baseline (0.000608 with r=0.296 > 0.0000293 ±0.000186 with r=0.00606 ±0.0394).
Small-World	**Small-world network but with longer paths** higher clustering coefficient (0.231 >> 0.0109 ±0.00511) but longer avg. path length (4.38 > 3.8 ±0.899) than random.	**Strong small-world features** higher clustering coefficient (0.695 > 0.338 ±0.00634) and comparable average path length (1.73 ≈ 1.66 ±0.00605) compared to random.	**Strong Small-world features** significantly higher clustering coefficient (0.752 >> 0.0536 ±0.00739) and comparable average path length (2.51 ≈ 2.61 ±0.0345) to random.

Although the typical range of the Log-Log slope for Scale-Free networks is between 2-3, our results are still indicative of that property, within the limit of the small dataset sample size. Logarithmic conclusions scale with the order-of-magnitude of the sample size with our ~100 points being at the lower threshold. The Indian Unicorn Startup network can be considered to being studied during a transitional phase, which has not yet fully matured into the classic scale-free structure, whose structure can only truly be observed across multiple orders of magnitude .

The steeper Log-Log slope for the Bipartite graph vs. the Investor-Investor graph can be attributed to the converse effect being bipartite has on path lengths — it introduces a more extended structure which limits the maximum degree of nodes making hub nodes more rare thus more pronounced.

These summary results conclusively demonstrate that Companies are not Scale-Free or Preferentially Attached, but Investors (and by extension the combined bipartite network) are Scale-Free and Preferentially Attached. All networks show evidence of high clustering with comparable path lengths to baseline random networks, indicating all are Small-World networks.

Conclusion

Our aim was to ascertain whether the evidence gathered through the network analysis of Indian Unicorn Startups either confirms or refutes predictions about demonstrating scale-free, preferential attachment, and small-world properties.

Assessment of Research Hypotheses

Hypothesis 1: Companies not Scale-Free or Preferentially Attached Largely confirmed — the evidence indicates a lack of scale-free structure within the unipartite companies network, as seen from the flatter slope in the log-log degree distribution compared to the random baseline. Furthermore, the absence of a strong preferential attachment pattern aligns with our initial prediction about the company network's nature.

Hypothesis 2: Investors are Scale-Free and Preferentially Attached Partially supported — we observed a moderate scale-free structure in the unipartite investors network and the bipartite company-investor network (which includes the investors), indicative of some hub formation, although not as pronounced as one might expect in a typical scale-free network. The preferential attachment was also evident but weak, suggesting conclusions with limited certainty, likely due to the small sample size of the dataset.

Hypothesis 3: All Studied Networks are Small-World Strongly confirmed — the high clustering coefficients and comparable average path lengths to random baselines across all networks analyzed unequivocally point to the presence of small-world characteristics.

Network Analysis Insights for Stakeholder Market Strategies

If these properties (all Small-World, Investors are Scale-Free & Preferentially Attached but Companies are not) are indicative of the broader VC market, this analysis can be generalised to provide strategy advice to different stakeholders, depending solely on their relative position in the network. Companies and investors could calculate similar metrics for their own business social networks, the then use them to inform their strategic decision-making the following ways

Investors: Incumbents

They are outliers in a scale-free network, which means they're connected to a large number of nodes, and benefit from Incumbency Bias thanks to Preferential Attachment. This position and their high Betweenness Centrality gives them preferential access to advantageous market information and deal-sourcing opportunities, letting the rich get richer and continue to outpace and out-scale their smaller competitors, by being more able to identify and raise larger funds for emerging opportunities across a broad range of sectors.

However because they are central bridging nodes, their position also requires them to be cautious of targeted vulnerabilities. They gain little upside from random additional co-investors, but should be selective to not take on too many risky investments with unproven companies or co-investors. So they should focus on leveraging their influence while ensuring robustness against targeted risks, such as breakdowns of key partnerships or failures in key sectors or portfolio companies.

Another advantage of scale is if like some large banks they become too central and make up significant proportions of market capital, during a financial downturn they could be deemed 'too big to fail' or more accurately 'too central to fail' (their failure would lead to a cascading collapse of the business network), thus historically such privileged nodes have received special treatment and favoritism from government assistance programs.

Investors: Disruptors

They are new entrants in a scale-free network, so will face challenges overcoming the Preferential Attachment that favours established players, by forming targeted alliances to increase their network Centrality. This way they can better integrate into the network and gain better access to deal flow.

The small-world nature of the network can allow them to quickly establish long-reaching connections. If they can successfully target and develop hard-to-source emerging and high-growth markets, staking out their own unique (high-value) network positions, this would increase their appeal as future co-investors to better-connected investors.

Companies: Incumbents

Operating in a non-scale-free but small-world network, these companies benefit from maintaining diverse connections. Their focus should be on broadening their network and continuously innovating

to stay relevant. Despite the even spread of connections, it's crucial for established companies to constantly seek new partnerships and ventures.

High-betweenness investors can open doors to expansive networks and resources, which is particularly beneficial for companies looking to scale or enter new markets. But because these companies are already well-established, it will be challenging for them to form additional new connections. However since the Companies' network is already highly connected, new connections from investors may not add significant marginal value, at which point managing .

Companies: Disruptors

Companies have stronger limits on how many connections they can form than investors have, so these companies should especially prioritise investors aligned with their market niche (network local cluster) to get on the radar of larger investors higher-up in the food-chain, whose investment would unlock more doors thanks to their more central network positioning.

The small-world nature of their network allows these companies to rapidly establish connections. Their strategy should focus on leveraging this high clustering to build a supportive network efficiently. It will be an uphill battle so unlikely for them to be able to beat incumbents at their own game and out-compete them in their target sector, so they can instead use their small scale and agility to seek out novel connections and new market opportunities (e.g. tech, digital transformation, AI disruption etc.). On a global scale Indian startups have a large total addressable market, but lower access to Venture Capital than e.g. in the USA, so many of the successful Unicorn companies have in fact adopted this targeted strategy.

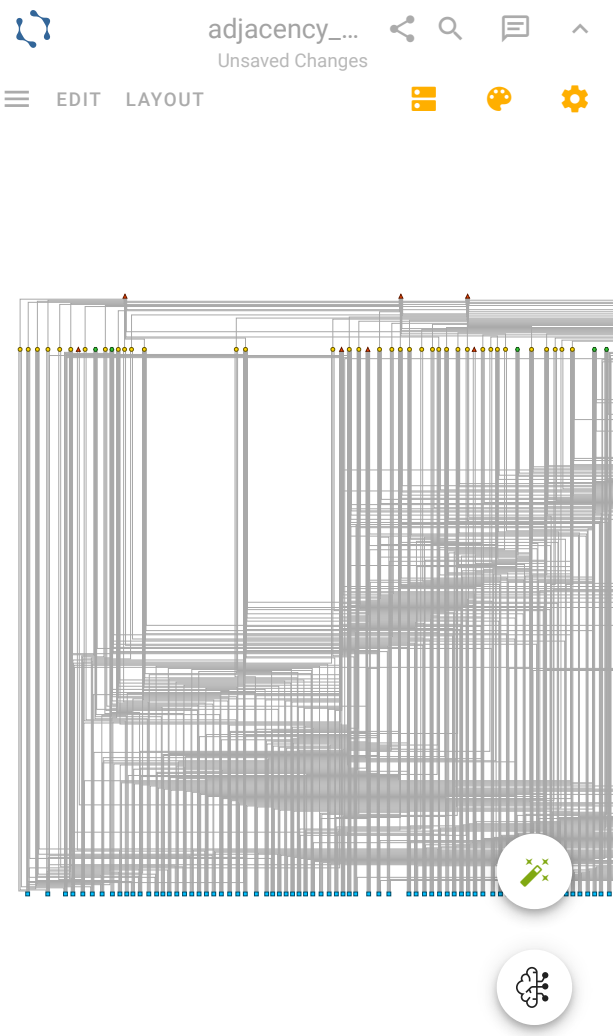
Review & Further Work

This investigation's primary technical challenges were the AI-enhanced web-scraping processing un-structured text for ~150 individual investors, and the challenge of organising 109 distinct graphs (with detailed tooltips) into a performant fully-interactive format.

The difference between the company and investor networks is strongly influenced by the sampling methodology - companies' investors were manually selected (with no documentation on the selection criteria), and the only investments observed were from the Unicorn sub-sample of the Indian sample of the global investment market, despite the majority of investors being from outside of India and likely investing outside of India. Thus the robustness of results would be improved by using a larger dataset to provide more consistent context and market coverage. Additionally it would be informative to test the robustness of the network (and confirm the scale-free property of being vulnerable to targeted disruptions) by comparing network metrics after selectively removing high-degree nodes (as is illustrated in [SM20]), but this is limited by the already small sample size of the network to begin with.

Likewise, community detection using clustering would be a worthwhile topic (correlating groupings within the network with groupings in their corresponding sector or region metadata), but with the current absence of any authoritative ground-truth data about causal mechanisms, attempting to recover meta-data is prone to over-fitting and methodologically flawed thus ill-advised [PL17].

One category of methods that could be employed are Graph Neural Networks (e.g. [CK22]), can in a sense be trained to compress connectivity and metadata features into an abstract embedding. Alternately the network could be enriched with sector and location metadata, enabling it to be treated as a knowledge graph.



**Interactive hybrid network/knowledge graph of the bipartite network with blue square (C) companies, yellow circle (I) investors (One-Time investors binned regionally),
and enriched with red triangle (L) Location, and green oblong (S) Sector nodes. Using a hierarchical display
(press 'Load File' to load the dataset, and minimise the displayed menus for larger view, change the display layout to 'Organic' for a density-based display)**

Finally, if we could train an algorithm to predict which companies a given investor is most likely to invest in to a sufficient degree of accuracy, this could reveal which features (extrinsic, or intrinsic) are most useful for helping to predict the future. The value of such a model would not only from being able to make predictions, but by using explainable ML techniques (e.g. SHAP [LU18] within XGBoost [CG16]) it could quantify the feature importance contributed by different elements.

Bibliography

- [ES01] English Standard Version Anglicised. (2001). *The Holy Bible.* Crossway Bibles, a division of Good News Publishers. Matthew 25:29 ESVUK.
- [WS98] Watts, Duncan J., & Strogatz, Steven H. (1998). *Collective dynamics of 'small-world' networks.* Nature, 393(6684), 440–442. https://doi.org/10.1038/30918
- [FS23] Fahlenbrach, Rüdiger & Stulz, René M. (2023). *Unicorns.* In The Palgrave Encyclopedia of Private Equity. Springer International Publishing. https://doi.org/10.1007/978-3-030-38738-9_130-1
- [YV17] Yan, Gang, et al. (2017). *Network control principles predict neuron function in the Caenorhabditis elegans connectome.* Nature, 550(7677), 519–523. https://doi.org/10.1038/nature24056
- [EG20] Evtushenko, Anna & Gastner, Michael T. (2020). *Beyond Fortune 500: Women in a Global Network of Directors.* In Complex Networks and Their Applications VIII. Springer International Publishing, 586–598.
- [LN23] Lutz, Eva & Nörthemann, Antonia. (2023). *Venture Capital Syndication.* In The Palgrave Encyclopedia of Private Equity. Springer International Publishing. https://doi.org/10.1007/978-3-030-38738-9_34-1
- [SC23] Siming, Linus. (2023). *Employment Networks in Private Equity.* In The Palgrave Encyclopedia of Private Equity. Springer International Publishing. https://doi.org/10.1007/978-3-030-38738-9_36-2
- [GM23] Garfinkel, Jon A., et al. (2023). *Alumni Networks in Venture Capital Financing.* SMU Cox School of Business Research Paper No. 21-17. https://ssrn.com/abstract=3970128
- [BA16] Barabási, Albert-László. (2016). *Network Science.* Cambridge University Press. https://books.google.fr/books?id=iLtGDQAAQBAJ
- [RP07] Robins, Garry, et al. (2007). *An introduction to exponential random graph (p*) models for social networks.* Social Networks, 29(2), 173–191. https://doi.org/10.1016/j.socnet.2006.08.002
- [GJ16] Galle, Caleb & Jespersen, Kristjan. (2016). *Transnational Markets for Sustainable Development Governance: The Case of REDD+.* World Development, 86, 79–94. https://doi.org/10.1016/j.worlddev.2016.06.009
- [LK04] Liben-Nowell, David & Kleinberg, Jon. (2004). *The Link Prediction Problem for Social Networks.* Proceedings of the Twelfth International Conference on Information and Knowledge Management, 556–559. https://doi.org/10.1145/956863.956972
- [NE01] Newman, M. E. J. (2001). *Clustering and preferential attachment in growing networks.* Physical Review E, 64(2). https://doi.org/10.1103/PhysRevE.64.025102
- [SM20] Saqr, Mohammed, et al. (2020). *Robustness and rich clubs in collaborative learning groups: a learning analytics study using network science.* Scientific Reports, 10(1), 14445. https://doi.org/10.1038/s41598-020-71483-z
- [PL17] Peel, Leto, et al. (2017). *The ground truth about metadata and community detection in networks.* Science Advances, 3(5). https://doi.org/10.1126/sciadv.1602548
- [CK22] Cosmo, Luca, et al. (2022). *Latent-Graph Learning for Disease Prediction.* In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, 643–653. https://doi.org/10.1007/978-3-030-59713-9_62
- [LU18] Lundberg, Scott. (2023). *Census Income Classification with XGBoost.* SHAP Documentation. https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/Census%20income%20classification%20with%20XGB
- [CG16] Chen, Tianqi & Guestrin, Carlos. (2016). *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. http://doi.acm.org/10.1145/2939672.2939785

Appendix I: Final Enriched Features List

This table contains the final feature set, combining binned and enriched sources. The only features used for the network analysis ended up being Names and 'company_investor1-7'.

Engineered Feature Name	Description
company_id	Unique identifier for each Indian unicorn startup
company_name	Name of the Indian unicorn startup
company_sector	Industry sector of the startup
company_valuation_unicorn_entry	Valuation at the time of becoming a unicorn ($B)
company_valuation_present	Current valuation of the company ($B)
company_date_unicorn_entry	Year the company became a unicorn
company_location_city_india	Indian city where the startup is headquartered
company_investor_all	All investors in the startup
company_location_city_international	International city of operation
company_location_latitude	Latitude of the company's location
company_location_longitude	Longitude of the company's location
company_sector_binned	Categorized industry sector
company_investor1-7	Investor Name
investor_name	Name of the investor
investor_description_pitchbook	Description of the investor from PitchBook
investor_startyear	Year the investor started
investor_website_url	Website URL of the investor
investor_location_country	Country where the investor is located
investor_location_city	City where the investor is located
investor_location_fullmailingaddress	Full mailing address of the investor
investor_location_latitude	Latitude of the investor's location
investor_location_longitude	Longitude of the investor's location
investor_global_region	Global region where the investor operates