Information Flow Simulation in an Investor Network

Janelle Gloriane R. De Vera, Felipe D. Garcia Jr., Bernadette M. Obiso, James Patrick T. Verdan
Aboitiz School of Innovation, Technology and Entrepreneurship, Asian Institute of Management
jdevera.msds2023@aim.edu, fgarcia.msds2023@aim.edu, bobiso.msds2023@aim.edu, jverdan.msds2023@aim.edu

Abstract

Using an SI Simulation on the investor network, an information flow simulation was conducted to determine the ideal set of investors startup companies should inform first. Different scenarios were conducted, and the spread of information was measured using two metrics: spread and relevance metric. The scenario that showed better information dissemination at the early simulation stages was when the initial investors were selected from the top eigenvector centrality of the same industry as the startup company. At t=6, this top scenario has 12 and 6 more informed investors than the random and top valuation scenarios.

Keywords: Network Science, Bipartite, Eigenvector Centrality, Susceptible-Informed Simulation, Investor Network

1. Introduction

The term "Unicorn" is used in the venture capital industry to refer to a private startup with a valuation of over $1 billion.[1] Companies such as Canva and Uber are famously known as Unicorn Startup companies. However, it is not easy for Startup companies to reach this valuation. They had to go through a series of funding stages that could take at least seven months per funding stage and still end up with a less desirable outcome. Particularly during the 3rd stage of funding, called Series A, where venture capitalist investments kick off, and company shares are exchanged for capital, setting the stage for significant growth and development. These investors not only fund the company but also offer mentorship, guidance, and valuable networking opportunities that could contribute to the growth and valuation of the business. As such, startups must enhance the quality and reach of their network to find the right investors that could fast-track their growth.

This study explores the possible scenarios through which a Startup company can boost its network quality and reach of investors.

Similar studies, such as a paper by Christian Esposito et al., entitled "Venture capital investments through the lens of network and functional data analysis," used different network centrality measures to capture the role of early investments in the firm's success. The results, which are robust to different specifications, suggest that success has a strong positive association with the firm's and its large investors' centrality measures. On the other hand, it has a weaker but still detectable association with centrality measures of small investors and features describing firms as knowledge bridges. Finally, based on their analyses, success is not associated with firms' and investors' spreading power (harmonic centrality), nor with the tightness of investors' community (clustering coefficient) and spreading ability (VoteRank). [3]

In another paper by Moreno Bonaventura et al., they look at large-scale online data to construct and analyze the time-varying worldwide network of professional relationships among start-ups. The nodes of this network represent companies, while the links model the flow of employees and the associated transfer of know-how across companies. It uses network centrality measures to assess the likelihood of a start-up's long-term positive economic performance in its early stages. They found that the start-up network has predictive power, and using network centrality can provide valuable recommendations, sometimes doubling the current state-of-the-art performance of venture capital funds. The network-based approach supports the theory that the position of a start-up within its ecosystem is relevant for its future success. At the same time, it offers an effective complement to the labor-intensive screening processes of venture capital firms. [4]

This study's contribution is applying an SI Simulation on the investor network to simulate the flow of information based on different scenarios to determine the ideal set of initial investors startup companies should inform first.

2. Data Description

The data for this study was sourced from Kaggle [5], which contained information on Unicorn Startup companies such as their valuation, date started, base country, city location, industry, and their list of associated investors, which is limited to a maximum of four investors per startup company.

The list of investors comprised branch companies, and these entities were considered as a single unit to get the unique list of investors from the data. This study includes data cleaning. The final dataset has 1,165 unicorn startup companies, including ByteDance, SpaceX, SHEIN, Stripe, and Canva, as the top five companies with the highest valuation and 1,212 corresponding investors containing prominent companies such as Google and Sequoia Capital.

3. Methodology

3.1 Data Preprocessing

The study's methodology is initialized by cleaning the investor column of the dataset. Two main processes are done to ensure that the dataset is prepared for the succeeding steps. One is by standardizing the names of the investors by looking at typographical errors in the dataset. This ensured that no nodes or investors were duplicated due to this error. The next process is to look into each unique investor, check if they are of the same company, and combine them into a single entity for the study. For example, "Sequoia Capital", "Sequoia Capital China", and "Sequoia Capital India" are unified into being "Sequoia Capital". This is based on the assumption that if an investor is part of a conglomerate, the information will surely be spread among its subsidiaries. On the other hand, we retained as two separate investors those with similar names but are in no way affiliated with one another.

3.2 Network Construction

The network is constructed by creating two nodes: startup company and investor. When an investor invested in that startup company, a connection links these two nodes. The network is designed to be an unweighted graph in which all the edges between nodes are considered to have equal importance or weight. In other words, no numerical values are associated with the edges to indicate varying degrees of importance or distance

between nodes. The resulting network is shown below.

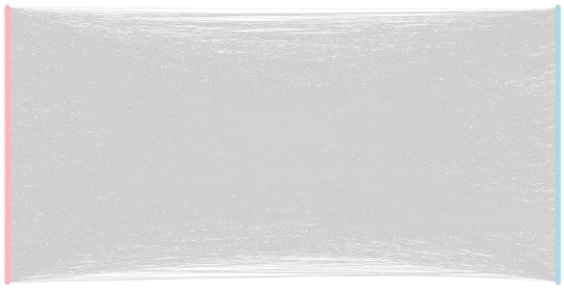

Figure 01. Investor-Startup Bipartite Network. The relationship between the startup company (blue node) and investor (pink node) is projected in a Bipartite Network. This network contains 2,374 nodes and 3,327 edges. The relationship is only between investor and startup, and there is no edge connecting investor to investor or startup to startup.

From this, the investor projection network was then generated. It shows the connection of investors to other investors if they invested in a common startup. From this, the investor projection network was then generated. It shows the connection of investors to other investors if they invested in a common startup. In this study, all simulations will only focus on the investor projection network.

3.3 Centrality Measure

Exploratory Network Analysis (ENA) is performed on the investor projection network. Here, the different network characteristics are analyzed to uncover the initial insights. This network statistics includes the nodes' minimum, maximum, and average degree, which relates to the number of edges connected to that node. Another is the clustering coefficient, which measures how nodes in a network cluster together. Lastly is the centrality measure, which focuses on Eigenvector centrality. Eigenvector centrality measures a node's influence based on its connections to other highly connected nodes.[6] Their centrality extends to other well-connected companies in the network, enhancing their overall influence.

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in V} a_{v,t} x_t \qquad (1)$$

Where:
x_v = eigenvector centrality score of node v
x_t = centrality score of node t
$a_{v,t}$ = adjacency matrix of value between node v and t
λ = constant (eigenvalue)
$M(v)$ = set of neighbors of v

3.4 Susceptible Informed Simulation

The study simulated information dissemination through Susceptible-Infected (SI) simulations, where *Infected* (*I)* is modified to be informed instead of infected. This is to pertain that information, instead of disease, is spread. The informed investors are those who were informed about the startup company; the susceptible are the opposite. Using an initial set of informed investors, the information is spread through the network with a probability (β). To change the status of susceptible investors, it must meet two conditions. First, the susceptible investor is connected to an informed investor. Second, the susceptible and informed investors are part of the same industry as the startup company. If the susceptible or informed investors have previously invested in the same industry as the startup company, the β = 0.10; otherwise, β = 0.01. The value was chosen conservatively, ensuring a higher probability was given to a situation with expected information transfer. [7]

3.5 Simulation Scenarios

In this paper, the researchers compare three different scenarios for the simulation. Each scenario has the same number of initial investors from the same industry as the startup company. The selection of these initial investors is different from each other. In scenario 1, the initially informed investors

were picked randomly, while in scenario 2, it was selected from the investors of the top startup company in terms of valuation. The last scenario determines the initial set of investors by selecting the top eigenvector centrality nodes. Scenario 3 simulates the effect of knowing the most influential investors and informing them first. The simulation was run at 20 timesteps and repeated five times to get average simulation results.

3.6 Metrics

The spread of information was measured using two metrics to compare the different scenarios. A spread metric was determined, which states how the information about the startup was spread in the investor network. This is represented by the mathematical equation shown below.

$$Spr_i = \frac{I_i}{T} \qquad (2)$$

Where:
Spr_i = spread metric at time step i
I_i = total number of informed investors at time step i
T = total number of investors in the network

Another metric used is the relevance metric. This metric measures how well the information spreads among the investors in the same industry as the startup company. It is defined as shown below.

$$Rev_i = \frac{In_i}{T} \qquad (3)$$

Where:
Rev_i = relevance metric at time step i
In_i = total number of informed investors belonging to the same industry as the startup at time step i
T = total number of investors in the network

4. Results and Discussion

4.1 Network Results

The investor projection network of the dataset resulted in 1,212 nodes, corresponding to the number of unique investors in the dataset. The number of edges is 3,010, which shows how connected the investors are to each other. With the dataset's four investor limitation, the network's resulting average degree is 2.29. This also shows that there are companies that have less than four investors related to them. In terms of the degree ranges, some nodes are not connected to other nodes, which could be interpreted that the company is the sole investor in that startup. The maximum degree of 145 indicates that some investors have extremely high influence, given the number of neighbors in their network. A clustering coefficient of 0.64 suggests that the network exhibits a relatively high level of clustering. This indicates that a significant portion of the nodes in the network have connections to other nodes within their local neighborhoods. This could imply that groups of investors tend to invest together or have common investments in an investor network.

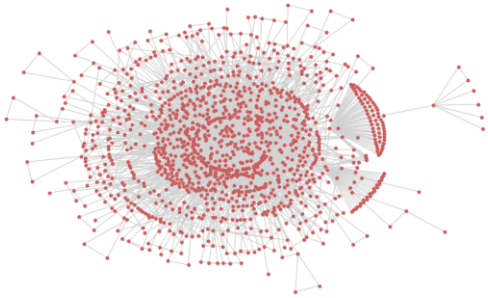

Figure 02. Investor Projection Network. The plot shows the relationship between the investors in which they are connected if they have a common startup company in which they are investing.

Exploring the eigenvector centrality of the investor projection network, the average value of the centrality is 0.01158, suggesting a relatively low score across the network.

A higher eigenvector centrality value indicates that a node is more central or influential than other network nodes. Conversely, a lower eigenvector centrality value suggests that a node is less central or influential. It implies that, on average, these nodes do not have strong connections to other highly central nodes in the network. This could mean that the network is decentralized, with many nodes having similar levels of influence and centrality. This can also mean that given the number of nodes in the model, there can be critical nodes where the eigenvector is extremely high and should be considered in the simulation. To illustrate this, the heatmap of the top 100 investors in terms of eigenvector centrality is projected.

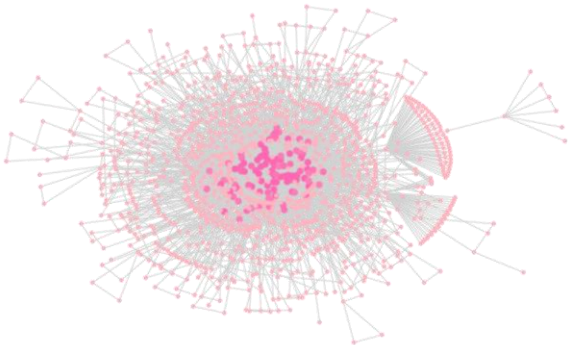

Figure 03. Eigencentrality Heatmap of the Investor Projection Network. The graph visualizes which points have the highest eigencentrality value and how they are connected.

The investors with the top 10 eigenvector centrality scores are extracted as they will be critical in the simulation by being one of the initial target investors in the modeling part of the study. It can be observed that Sequoia Capital has the top score by a considerable margin. A node with relatively higher eigenvector centrality than others in a network is exceptionally central and influential due to its extensive connections, a potential link to other highly central nodes, or strategic positioning. This high eigenvector centrality often signifies its pivotal role in controlling information or influence flow within the network, making it a crucial player in its dynamics.

4.2 Simulation Results

This paper simulates the information spread for a new Artificial Intelligence (AI) startup company. The three scenarios are established, and the results are compared.

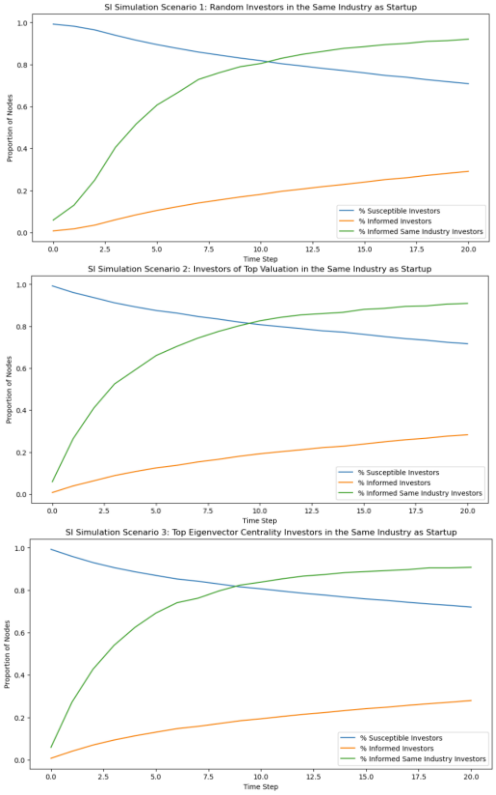

Figure 04. Simulation results of three scenarios tracking the percentage informed. All three scenarios display the same plot behavior with a difference in the value. The susceptible investors decrease over time as the informed investors increase. At the t=20, these two values do not intersect yet but experience gradual change. The informed investors in the same industry increase rapidly at timestep 0 to 10 but plateau afterward. This behavior occurs because it gets harder to spread the information to the remaining susceptible since these investors are only connected to a few.

Simulation results show how the network population changes its status from susceptible to informed. The diagram across all scenarios acts similarly, with the difference only in value. This difference is further explored in the metric comparison.

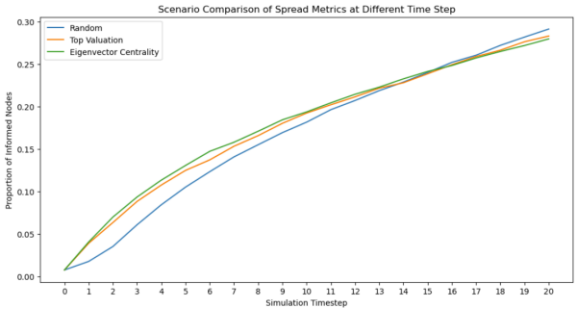

Figure 05. Spread metric comparison of different scenarios. Using eigenvector centrality and top valuation spreads the information faster at an earlier period ($0 \leq t < 16$) than random by a margin. This difference in spread decreases over time as the random scenario catches up and closes the gap. At $t = 16$, random spreads more than the other two scenarios.

Using the scenarios based on the top eigenvector centrality and the top valuation, investors tend to spread the information across all investors faster during the early stages. On the other hand, the random scenario tends to be better after a long period. Time steps in this simulation are not synonymous with a day or month. Instead, it represents a change in investors' information knowledge after some time, which may take months to years, depending on the interaction between connected investors. For startup companies, it is crucial to spread the information to as many investors as fast as possible to get funds for the company early on. These results show the information dissemination for all investors, regardless of the industry. To have a more focused industry perspective, we check the relevance metric.

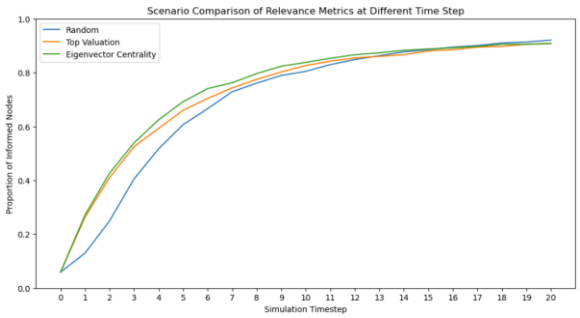

Figure 06. Relevance metric comparison of different scenarios. Using eigenvector centrality and top valuation scenarios, the information spreads to industry investors faster at an earlier period ($0 \leq t < 16$) than random by a large margin. This difference in spread decreases over time as the random scenario catches up and closes the gap. At $t = 16$, random spreads more than the other two scenarios.

The relevance metric shows a similar behavior as the spread metric, with eigenvector centrality and top valuation showing a faster spread at the start. However, the random scenario catches up towards the end. The difference between the relevance and spread metrics is the larger gap between the random and the other two scenarios at the early stages. This represents a larger difference in the number of informed investors in the same industry as the startup.

The eigenvector centrality scenario leads by a few percentages in the relevance metric against the top valuation scenario. Choosing the top eigenvector centrality scenario is better demonstrated when comparing the absolute value of the results of the relevance metric. At t=6, the eigenvector centrality scenario has informed 125 industry investors (74% of the total 169 industry investors), while the top valuation scenario has informed 119 industry investors (70%). The additional six (6) investors could mean a difference, having informed the investor that might fund the startup company.

5. Conclusion

The scenario where the initial investors are chosen based on eigenvector centrality tends to have better information spread at the early stages ($t < 16$) of simulation. It outperformed the random scenario by having 12 more informed investors and six (6) more than the top valuation scenario. This simulation proves that when a startup company wants to spread the information, selecting the initial informed investors is crucial as it will affect the information spread throughout the investor network. By choosing the investors with the highest eigenvector centrality, the startup company can spread the information faster in the early stages, which is crucial to getting the word out about the startup.

6. Recommendation

The authors recommend having a more tailor-fitted simulation that resembles investor interaction. As there exist investor groups that could make the flow of information faster. Its corresponding simulation would have a higher probability β than the current study's assumption as the information within this investor group would be disseminated faster. Conducting a study to determine this probability on actual information flow scenarios in the startup community is recommended.

In addition, the study can be translated to non-startup company-investor networks as this study can still benefit companies looking for investors.

7. References

[1] Chen, J. (2022, May 31). Unicorn Definition. Investopedia. https://www.investopedia.com/terms/u/unicorn.asp

[2] The 8 Stages of Startup Funding. (n.d.). Indeed Career Guide. https://www.indeed.com/career-advice/career-development/startup-funding-stages

[3] Esposito, C., et al. (2022, June 27). Venture capital investments through the lens of network and functional data analysis. Applied Network Science. https://appliednetsci.springeropen.com/articles/10.1007/s41109-022-00482-y

[4] Bonaventura, M., et al. (2020, January 15). Predicting success in the worldwide start-up network. Scientific Reports. https://www.nature.com/articles/s41598-019-57209-w

[5] Unicorn Startups. (n.d.). www.kaggle.com. https://www.kaggle.com/datasets/ramjasmaurya/unicorn-startups

[6] Iacobucci, D., et al. (n.d.) Eigenvector Centrality: Illustrations Supporting the Utility of Extracting More Than One Eigenvector to Obtain Additional Insights into Networks and Interdependent Structures. Journal of Social Structure. https://www.cmu.edu/joss/content/articles/volume18/IacobucciMcBridePopovich2017.pdf

[7] Hoffmann A. & Broekhuizen T. (2009, January 20) Susceptibility to and impact of interpersonal influence in an investment context. Springer Link. https://link.springer.com/article/10.1007/s11747-008-0128-7

8. Appendix

List of investors for scenario 2 (top valuation) and 3 (eigenvector centrality). Scenario 1 investors change every run as they are selected randomly.

Table 01. List of initial ten (10) investors for scenario 2. Investors who invested in the same industry as the startup, such as Artificial Intelligence, and also invest in the top valuated startups.

Investors	Top Valuation (Billion USD)
Sequoia Capital	ByteDance ($ 140)
SoftBank Group	ByteDance ($ 140)
Sina Weibo	ByteDance ($ 140)
SIG Asia Investments	ByteDance ($ 140)
Founders Fund	SpaceX ($ 127)
Tiger Global	SHEIN ($ 100)
Khosla Ventures	Stripe ($ 95)
capitalG	Stripe ($ 95)
DST Global	Checkout.com ($ 40)
Insight Partners	Checkout.com ($ 40)

Table 02. List of initial ten (10) investors for scenario 3. Investors in the same industry as startup with the highest eigenvector centrality.

Investors	Eigenvector Centrality Score
Sequoia Capital	0.661
Tiger Global	0.096
Accel	0.091
Tencent	0.089
Insight Partners	0.068
SoftBank Group	0.056
Andreessen Horowitz	0.053
Lightspeed Venture Capital	0.052
Index Ventures	0.051
Matrix Partners	0.050

There are three common investors between these two scenarios, as they are part of the investors with top valuation and high eigenvector centrality.