Into the Billion Dollars: Geolocation Influence and Link Prediction Analysis over Unicorn Startups.

Reem Aleithan
ID: 216051765
Login: reem1100
Lassonde School of Engineering
Toronto, Ontario
reemaleithan@gmail.com

Mohammadreza Karimi
ID: 215840853
Login: r3za
Lassonde School of Engineering
Toronto, Ontario
r3za@my.yorku.ca

Mahmoudreza Eskandarijam
ID: 214420327
Login: rezajam
Lassonde School of Engineering
Toronto, Ontario
rezajam@my.yorku.ca

ChengXiang(Leo) Gong
ID: 214321616
Login: q5787882
Lassonde School of Engineering
Toronto, Ontario
q5787882@my.yorku.ca

INTRODUCTION/MOTIVATION

The ever-expanding world of investors and startups has gone through many changes over the years and yet continues to grow over different industries, regardless of economic fluctuations. As a result of these continuing investments, some startups gain a high value in the market, and eventually become known as Unicorn startups (worth over $1 billion$ of the market share). This continuous phenomenon brought interest and vision to us in the field of network analysis.

After intensive research of how investments are made and what influences certain investors to invest in some type of industry over the other, we decided to take this study further and understand how some investments influence others as well as predicting future possible investments. For the analysis, we used company-investor related data from Crunchbase [1]. However, in order to provide a better-focused analysis, we decided to limit the scope of the data to only Unicorn Startups and their top five investors over different contents and between 2017 and 2020.

With such strong dataset, we plan to understand how the investment network behaved over the past 4 years. To expand, with the help of extra research about social influence and provided the location of each investor, we are interested in understanding how certain investors socially influenced others in their vicinity to invest in similar companies.

To extend, we also used the understanding of the current investment network to predict future possible investments.

1 PROBLEM DEFINITION

1.1 Definitions

The data we collected is from the recent released information from Crunchbase (an online data platform for research over private and public companies). The specific dataset we focused our analysis on is unicorn startup companies and their corresponding top five investors.

Using this dataset, we wanted to see if an investor within some location influenced other investments within that same content the investor is based in. Furthermore, we hope to analyse if there is a cross-industry investment influence within the same content, i.e. if an investor based in China made an investment in the software industry influenced other investments within Asia but in a different industry (e.g. artificial intelligence, or healthcare).

As a further expansion, provided the top five most funded investors for each company in the network, we wanted to see what will be the potential companies for investment, based on their own circle of trust, their common neighbor investors and other similar investors, for 2020.

1.1.1 ***Company:***

- **Unicorn Startups:** Unicorn is a term in venture capital industry that describes a startup company, that has a value of over $ 1 billion [4].

In our collected data-set, we stored information about unicorn companies with last funding round ranging from 2017 until October 28, 2020. Such information included: their valuations, the industries they belong to, their total equity and many more explained below.

1.1.2 ***Investor:***

- **Definition:** A person or an entity (such as a company) that commits capital (i.e. gives money to some company) expecting to get some financial returns.[3]

For this project we decided to consider the given top 5 investors from the Crunchbase.

In general, looking at the data acquired, we define a company as the rows in which they have investors for. Majority of such companies and the basis of our data research is the unicorn startups but in order to make a better analysis and more interesting network, we also considered some of the investors with their own investors as companies like ***"Tencent Holdings"***, as well as some of the companies that invested in others as investors.

Project EECS4414, Nov 7, 2020, Toronto, Ontario, Canada
2018. ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1.1.3 *Investors' Geolocation Influence Related Definitions:*

- **Active node:** Investor that has a timestamp for a certain activity.
- **Friends:** Investors within a certain geolocational distance from each other.

1.2 Limitations:

For the Company-investor connections, two types of networks have been created so far, undirected and directed. For our prediction model we decided to only use the undirected graph, because our prediction algorithm only cares about whether two nodes are connected and doesn't care about the direction of the edges.

In addition, throughout the process of getting the link prediction via PageRank algorithm, the network required ***max-iterations*** to be huge and therefore in some cases not even reaching the epsilon convergence, hence we decided to increase the default epsilon value in NetworkX slightly.

$$\sum_i |r_i^{t+1} - r_i^t| < \epsilon \tag{1}$$

2 RELATED WORK

2.1 Investors' Geolocation Influence

Inspired by the "Individual behavior and social influence in online social systems" paper, essentially the project extends to see if there existed certain investors who with their investments influenced their surrounding investors to take similar actions.

In general, the paper explains that for certain problems involving a network of nodes, with existing attribute of timestamp, it is possible to figure out the existence of social influence if over the time that a similar activity starts from(i.e in their case, geotagging their picture on Flickr), other "friends" of that node also partake into the same activity after. This is how social influence is defined.

2.2 Link Prediction

The funding of investment connects the whole investment environment and they always lead us to the formation of complicated information networks. In the course of our research, we read some relevant research reports to increase our understanding of the study we are doing. The most relevant report we found to our research is *Venture Capital Investment Networks: Creation and Analysis*[6].

It is a research report from Stanford University in 2018. In their report, they focused on creating investment networks for venture capital, analyzed investment status through information obtained from data they found on Crunchbase from 2010 to year 2013. Instead of all the companies around the world, they paid attention to early-stage, start-up companies within networks they created. They got a sample of 18000 start-ups, nearly 4700 acquisitions and 52000 investment events as their data.

In their work, they used data: Companies, Rounds, Investments and Acquisitions to create Investors-to-Companies, Investors-to-Investors and Companies-to-Companies graphs and calculated Density, Diameter and Clustering Coefficent for each of the graph to get an overall idea for their data. By calculating degree centrality using $EigenventorCentrality^{(1)}$ for each graph they got top-5 types of investment from their data.

$$c_{eig}(x) = \frac{1}{\lambda} \sum_{y \to x} 2^{-n} c_{eig}(Y) \tag{2}$$

for comparing the weight between two companies with number of investors they share, they used $JaccardIndex^{(2)}$, which is defined as follows:

$$JA(i,j) = \frac{|\Gamma_i \cap \Gamma_j|}{|\Gamma_i \cup \Gamma_j|} \tag{3}$$

By doing the community detection with $LouvainAlgorithm$ and $Node2VecClustering$ they concluded and created a useful network representation of 2013 data set from Crunchbase on start-up Companies.

Another closely related paper we found was *Link Prediction in Bipartite Venture Capital Investment Networks*[2]:

In this paper, the data was also retrieved from Crunchbase. However, unlike the previous paper mentioned above, it mainly focused on the Crunchbase Business graph which includes the following: relationships and interactions that occurred between 280k unique persons, 300k unique organizations, 150k investment rounds and 16k acquisitions.

The network they created in the research was a bipartite graph, separating investors from companies. The links in the graph are directed edges from the investors to the companies they invested in. As a contrast to our own work, our graphs used attributes to identify companies vs. investors vs. companies that are also investors as opposed to separating the nodes using a bipartite division.

As evaluation of their work, they seperated the data over 3 month intervals and checked whether their predicted linked were actually formed after the 3 months they performed the prediction over. In our work however, we used the entire dataset for the prediction and evaluated our results through a automated process that we'll describe below.

Finally, what differentiates our project from their work is the chosen algorithms to perform the link prediction. We used Salsa algorithm while they used several different ones such as Random Link Predictor, Preferential Attachment Link Prediction, and Weighted Preferential Attachment Link Prediction. Since our data only has top 5 investors and we want to use investors information for the link prediction, Salsa was the best option to perform this task since it calculates the circle of trust for each investor.

Because of the recent changes in the economic environment and the impact of the Pandemic on the investment environment, we decided to further study on those successful emerging companies on the basis of previous report with data set we get on Crunchbase, and see what has changed because of the Pandemic. In our research project, we are investigating unicorn Start-up companies' investment situation with their data records on Crunchbase from 2017 to 2020, which is more recent data compared to previous work. In terms of data analysis, we will create graphs like Companies-to-investors and investors-to-investors to analyze the data set we have, and with the data we got from Crunchbase, we will be able to carry out clustering, degree distribution and even community detection and link prediction. They have been shown effective in our information network study. By having the updated data, a more narrow research topic and investment recommendation function, we intend to have a more focused and updated research than the mentioned articles.

3 METHODOLOGY

3.1 Investors' Geolocation Influence

The way this paper [5] gets the social influence is that it first checks this influence from the actual graph and timestamps by counting the "active" nodes and counting the numbers of social influences, and finds the relation of social influenced nodes over active ones. And comparing this with the the null model(i.e the model with the same connections but shuffled timestamps), to see if their difference is positive (i.e difference of Real model and its Null model)

In general, the project seeks to find certain concepts in the mentioned paper, like the timestamps, friendships and relations of nodes, and timeline of the activity in order to achieve the similar analysis. Below shows a brief explanation of the approach and its parallel representations how it relates to investors, their geolocations, and the investment timestamps.

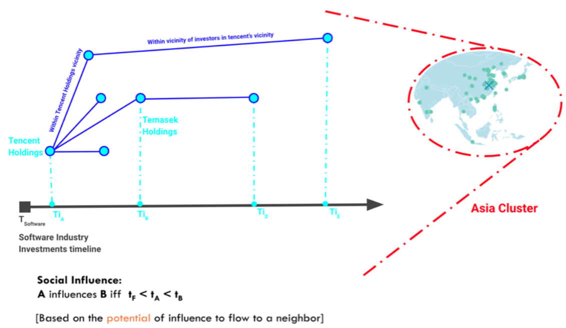


Figure 3.1.1: A brief glimpse of what the social influence of investors is being analyzed within one of the clusters, the Asia Cluster.

3.1.1 *Geo tagging and clustering.*

In order to achieve this first the headquarters geolocation tags of each investor and then, even with their different granularity, latitude and longitude was accumulated using "GeoPy" api And since The results were distributed almost everywhere around the globe, they were separated into different groups by applying clustering. The algorithm used is the K-means algorithm and the appropriate K number of clusters was determined to be K=3 by the Elbow method.

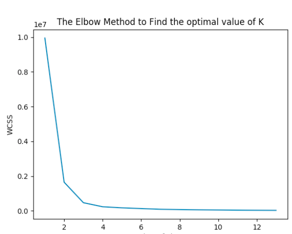


Figure 3.1.2: The results from the Elbow method.

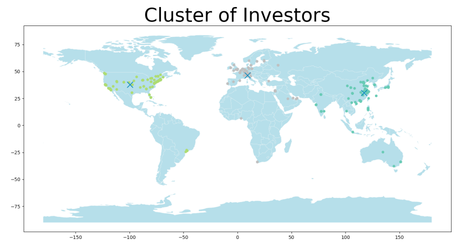


Figure 3.1.3: This was a correct step towards having more focused analysis of investors and their localities. As a matter of fact the three clusters simulated investors groups into three continents of Asia, Europe, and America more or less.

This section is heavily inspired by the paper mentioned [5], therefore there exists a concept similar to "most active geotaggers". So before diving into each "continent", to raise the chances of finding the influencing investors, the above mentioned "most important nodes" were selected for each cluster based on their common appearance in top 20 results of the mentioned algorithms. The results are as follows:

Most important investors based on clusters	
Cluster	**Most Important Investor**
Asia	Tencent Holdings
Europe	SoftBank Vision Fund
America	Goldman Sachs

Table 1: List of most important investors from common top 20 results of centrality algorithms mentioned

3.1.2 *Friends and friends of friends.*

Furthermore, with these results, for each cluster the distance of nodes to that cluster's most important node was determined and the ones with less than half of the average distance were called the investors within the vicinity or as we call them the "friends". Moreover, the friends who had other nodes within quarter of their average of distance were the "friends of friends".

So for instance for the cluster of Asia, all the distances to "Tencent Holdings" were determined and the average distance is 10460.104470197055 kilometres, and any node less than half of this are deemed as friends. Then these friends with nodes less than quarter of their average are the friends of friends. This resulted in the following directed graph:

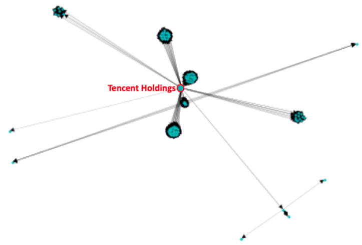


Figure 3.1.4: Directed graph of Asia cluster with "Tencent Holdings" as the most important investor.

3.1.3 Industries and timestamps of investments. Essentially the "Last funding date" for the company could summarize the decision of investment to be done by that time, for each investor. Also there were many investments in different companies of different industries. Having these two matters, the decision was to get, for each investor, its timestamps of investments for each industry, and then average those timestamps. As a result for each investor there exists the whereabouts of investments for each industry they invested in. For example some investors in the Asian's cluster the "Temasek Holdings" have made investments in Software around Feb 6, 2019, and so many more. And "Horizons Ventures" have made investments in the Internet around Sep 26, 2019.

At this point for each cluster each investor has information about their investments industries and the average timestamp for each of them. Clearly some did and some did not invest in certain industries and therefore based on top 5 mostly invested industries, the attributes of the network were set using networkx. For instance in the Asia's cluster, the top 5 most investments were done for Unicorn Startups in Software, Internet, E-Commerce, AI and Financial Services.

3.2 Link Prediction

3.2.1 Data Retrieval and Storage. As mentioned above, we used data from Crunchbase to create our graphs. However, the data we accessed from Crunchbase couldn't be downloaded in a desired format for our analysis. Therefore, we web-scraped it ourselves to be able to store it in an xlsx format.

The xlsx sheet stores the following information about each unicorn company: Location, No. (a unique id), Company Name, Description (of what the company does), Total Equity Funding, Valuation, Valuation Date, Industries, Top 5 Investors, Founders, Founded date, Funding Status, Last Funding Amount, Las Funding Date, Last Funding Type, Number of Acquisitions, CB Rank. The data includes 630 unique companies, 985 unique investors, and 18 unique companies that are also investors. For this part of the project, we were mainly interested in the company names, and the top 5 investors for each of them. Hence, we created a dictionary of company names as keys and their top 5 investors as the values for easy access when creating the graphs.

3.2.2 Graph Creation Using NetworkX. Using the data we stored as mentioned earlier, we created a company to investor graphs as follows:

- **Nodes:** we had three different kinds of nodes, 1) companies 2) investors, 3) companies that are also investors. we added an attribute value to each node and a color to distinguish its type from the 3 listed above.
- **Links:** the links between the nodes represents an investment between the company and the investor it's linked to.
- **Graph Direction:** we created two graphs, a directed (from the investor to the company it's investing in) and an undirected graph. We created the directed graph because it would make it easier to create different desired graphs later on (investor to investor graph, and company to company graph) for further analysis. While the undirected graph was use for the link prediction done in this project.

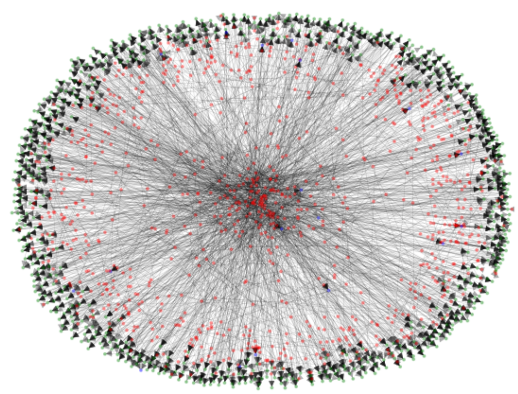


Figure 3.2.5: Visualization of the directed version of company-investor graph where the arrows go from investors to companies they invested in.

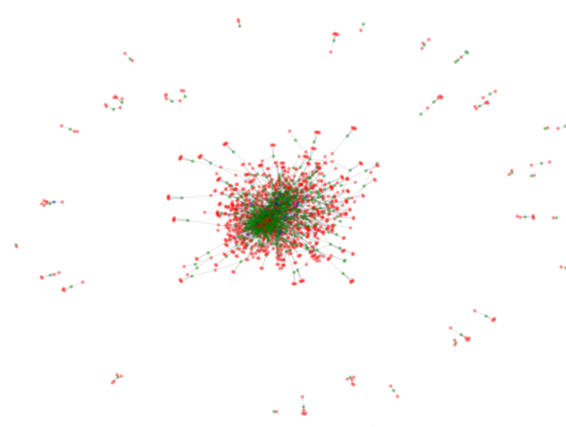


Figure 3.2.6: Visualization of the undirected version of company-investor graph where there is a connection between an investor and a company if the investor has invested in that company.

Figure 3.2.5 visualizes the directed version of the Company-Investor graph. The arrows go from the investor to the company they invested in. Figure 3.2.6 visualizes the undirected version of the Company-Investor graph. An investor and a company are connected together if the investor has invested in that company. In both of the graphs above, companies, investors, and company-investors are shown in green, red, and blue nodes respectively.

Also, figure 3.2.6 shows that most of the nodes in the network fit in the giant connected component and only few of companies and investors create a small network with internal connections.

3.2.3 Feature Extraction. We used Node2Vec to get multidimensional representations for each of the nodes in our graph. Node2vec returns a d dimensional feature vector for each node and it tries to minimize the cosine similarity between two nodes in this d dimensional space if the probability of reaching from one node to the other one is higher in a random walk. d is a hyper-parameter that you have to set in Node2Vec. In this project we set d to be 64. So after running Node2Vec, we had a 64 features for each node in our graph. But we can't really visualize a 64 dimensional node embedding. In order to see how our nodes are distributed in our multi dimensional space, we used Principal Component Analysis (PCA). PCA projects the data from a higher dimensional space, to a lower dimensional space and it does it by minimizing the projection error. We used PCA to project our 64 dimensional data into a 2d space.

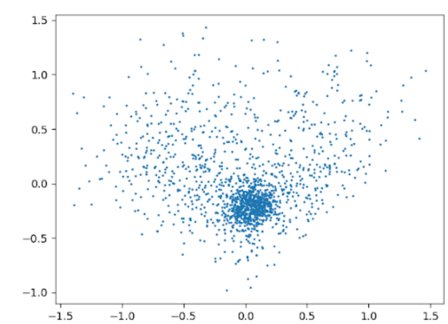


Figure 3.2.7: Projecting 64 dimensional node embedding into a 2d space using PCA.

Figure 3.2.7 shows how our features looked like when they were projected into a 2d space using PCA. In the figure we can see that there seem to be 3 main clusters in our nodes. We used K-Means to find 3 clusters based on our embedding. We ran K-Means on our 64 dimensional feature vectors and then mapped the results back to 2d using PCA.

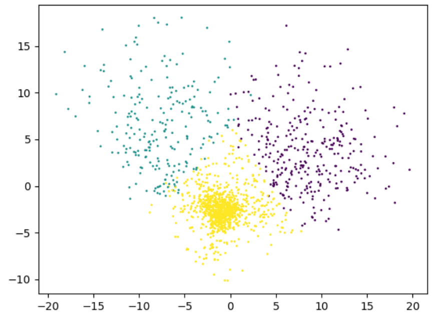


Figure 3.2.8: Finding 3 clusters using K-Means and projecting the results into a 2d space using PCA.

Figure 3.2.8 shows how our nodes were clustered. We wanted to know how this clustering look like in our graph, so we mapped the color of each cluster to our original graph.

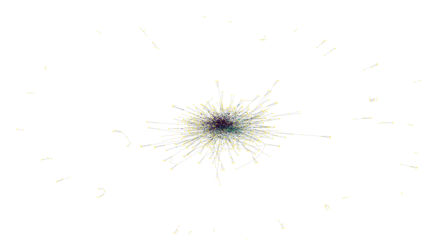


Figure 3.2.9: Mapping the clustering results back to the original graph.

Figure 4.1.17 shows how the graph looks like once we mapped the results of our K-Means clustering to the undirected graph. You can see that the green and purple clusters are more focused on the giant component of the graph, whereas the yellow cluster mostly includes nodes that form small disconnected subgraphs.

But we were interested in knowing what this clustering actually means. More importantly, we were interested in finding out whether it was related to the field/industry of companies. So for each cluster, we counted the number of companies in each industry, and created a histogram.

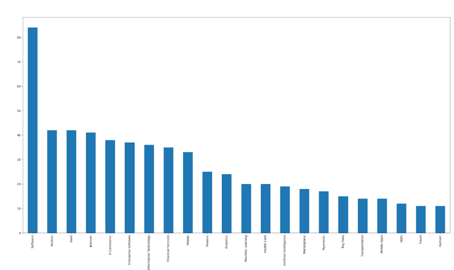


Figure 3.2.10: Industry histogram for companies in the purple cluster.

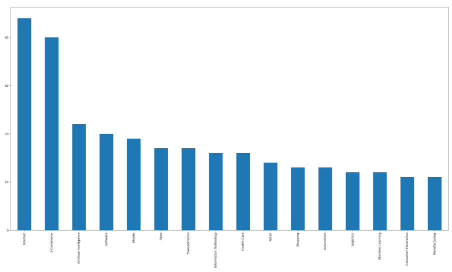


Figure 3.2.11: Industry histogram for companies in the green cluster.

Figure 3.2.10 shows the histogram for the purple cluster and figure 3.2.11 shows the histogram for the green cluster. We can see that in each cluster, there is one or two dominant industry. We didn't include the histogram from the yellow cluster, because it's mostly including small disconnected subgraphs which could include companies from all industries and no dominant industry could be observed in its histogram. But as we increased the number of clusters, the same behaviour could be observed. There was always a cluster which included the disconnected subgraphs and had a noisy histogram, but all the other clusters, had one or two dominant industry in their histogram. We tired K-Means with k being 5 and 7 and we got the same observation.

But what does this mean? We can interpret this observation as having a close connection between the field/industry a company works in, and the field/industry an investor is most interested to invest in. In other words, Investors, tend to invest in companies in certain industries. So a link prediction algorithm that works based on similarities between investors, will most probably perform well.

3.2.4 Link Prediction. Salsa link prediction algorithm works exactly as we wanted. It works based on the similarities of the nodes. We Used Salsa to create an investment suggestion system. In other words, given an investor, we are interested in knowing what companies should that investor invest in based on the behaviour of investors that are similar to them. To do this we used the Salsa algorithm:

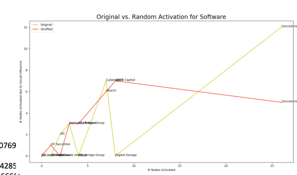
- First we need to find what investors are similar to the given investor. To do this, we used the Personalized PageRank algorithm. This can be achieved in NetworkX by running the builtin pagerank method and setting the probability of restarting in the given investor node to 1.0 and every other node to 0.0.
- Once we have the Personalized PageRank scores, we sort them and pick the nodes that are investors, and had the highest PageRank score. This creates the circle of trust of that investor, which is a set of investors that behaved similar to the given investor. In this project, we set the size of the circle of trust to be 7 because it leads to a relatively large sub graph. We refer to this circle of trust as Hubs.
- Once we found the circle of trust of the given company (hubs), we find the nodes that the Hubs are connected to. We refer to this set of nodes as Authorities. Then we create a sub graph which includes the nodes from both Hubs and Authorities.
- After we create the sub graph, we run the Hits Algorithm on it, which gives us an authority and a hub score for each node. When a node has a high authority score, it means that more nodes are connecting to it in the network and/or it's connections are coming form more important nodes in the network. So to find what companies the investor should invest in, we need to find the nodes of type company that have the highest authority scores. So we sort the list of authority scores for each node and pick the ones that their type attribute is company and are not connected to the given investor. The list of companies that we end up with will be the companies that are suggested to the investor for further investments.

4 EVALUATION

4.1 Investors' Geolocation Influence

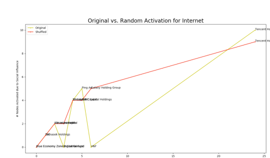
Eventually using, in each cluster, the directed graphs sourcing from the most important node with the mentioned attributes was created. Then in order to evaluate the social influence for the immediate nodes, the breadth first search traversal was used. This was then done over the rest of other nodes of the network as well, and after each traversal the results were compared with the null model. The following are the results from clusters:

The following are the some of the results of the influencing investors:



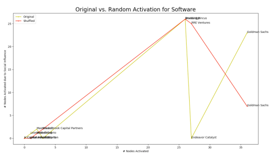
Influencing investors	Score
Tencent Holdings	0.269230769230769
Temasek Holdings	0.5
SoftBank	0.285714285714285
Everbright	0.666666666666666
Tiantu Capital	1.0
GIC	1.0
Steadview Capital	0.333333333333333
Macquarie Group	0.333333333333333
NTT Data	0.5714285714285714
China Merchants Capital	0.666666666666666
CyberAgent Capital	0.1428571428571429

Figure 4.1.12: The Asian investors influencing investments for Software industries.



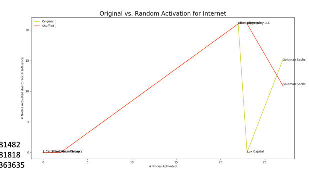
Influencing investors	Score
Tencent Holdings	0.0416666666666685
DST Global	0.25
CDH Investments	0.25
Tiantu Capital	0.6
Fortune Venture Capital	0.2
Elevation Capital	1.0
YF Capital	0.75
Ping An	0.19999999999999996
China Merchants Capital	0.2
China Merchants Bank	0.20000000000000007
CyberAgent Capital	1.0
Gaw Capital Partners	0.5

Figure 4.1.13: The Asian investors influencing investments for Internet industries.



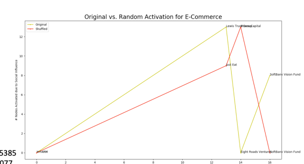
Influencing investors	Score
Goldman Sachs	0.4444444444444444
General Atlantic	0.4230769230769231
Morgan Stanley	0.5384615384615384
Kohlberg Kravis Roberts	0.6538461538461539
Mastercard0.5Insight Partners	0.6538461538461539
Leonardo DiCaprio	1.0
FJ Labs	0.038461538461538436
Neuberger Berman Group	0.15384615384615385
Access Industries	0.19230769230769224
Warburg Pincus	0.07692307692307687
GCP Capital Partners	0.07692307692307687
Arena Holdings	0.5769230769230769
Goldman Sachs Investment Partners	0.3846153846153847
Citigroup	0.4615384615384615
Coatue	0.23076923076923078

Figure 4.1.14: The American investors influencing investments for Software industries.



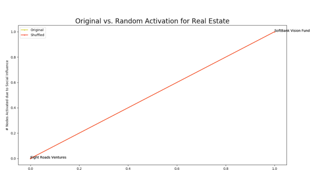
Influencing investors	Score
Goldman Sachs	0.1481481481481482
General Atlantic	0.6818181818181818
Morgan Stanley	0.13636363636363635
Kohlberg Kravis Roberts	0.6363636363636364
RRE Ventures	0.3636363636363637
Lead Edge Capital	0.5
Intonation Ventures	0.6363636363636364
Allen & Company LLC	0.7272727272727273
NBCUniversal	0.7727272727272727

Figure 4.1.15: The American investors influencing investments for Internet industries.



Influencing investors	Score
SoftBank Vision Fund	0.583
North	0.15384615384615385
Expand Venture	0.3076923076923077
Sprints Capital	0.3076923076923077
Just Eat	0.23076923076923084
DC Thomson Ventures	0.23076923076923073
Lewis Trust Group	0.8461538461538461
LocalGlobe	0.6923076923076923
Entrée Capital	0.6923076923076923

Figure 4.1.16: The European investors influencing investments for E-Commerce industries.



NOTHING

Figure 4.1.17: The European investors influencing investments for Real-Estate industries, in which no results were found

4.2 Link Prediction

In this section, we will first look at some basic metrics and then we see how we evaluated our link prediction algorithm.

4.2.1 Basic Metrics. Here, we will look into some of the basic metrics of our Company-Investor graph's Giant Component:

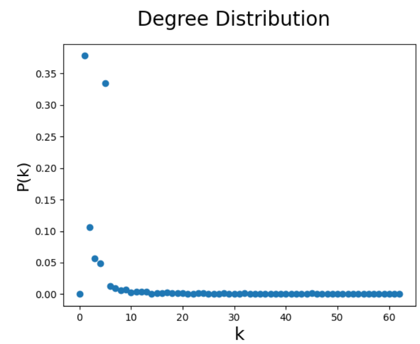


Figure 4.2.18: Degree Distribution of the Giant Component of the undirected graph.

Figure 4.2.18 shows the degree distribution of the graph. We can see that it is closer to a power law distribution.

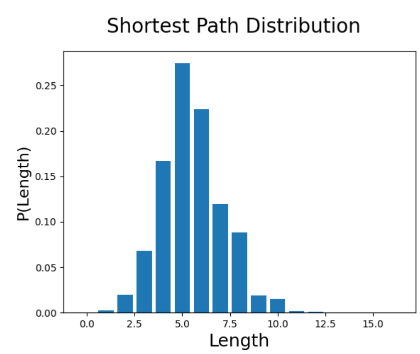


Figure 4.2.19: Shortest Path Distribution of the Giant Component of the undirected graph.

Figure 4.2.19 shows the shortest path distribution of the graph which has a binomial distribution. The average shortest path is 5.5 as noted in table 2.

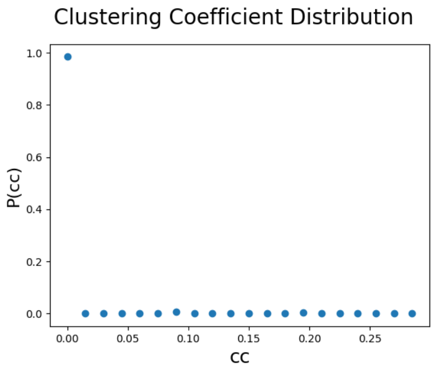


Figure 4.2.20: Clustering Coefficient Distribution of the Giant Component of the undirected graph.

Figure 4.2.20 shows the clustering coefficient distribution of the graph. We notice that most of the nodes have a very small clustering coefficient and this can be verified by checking the average clustering coefficient from table 2.

Network of Companies-to-Investors	
Metrics	**Undirected**
# of Companies	630
# of Investors	985
# of Companies-Investors	18
Density	0.0022
# of nodes of the whole graph	1597
# of edges of the whole graph	2830
# of nodes of the GCC	1457
# of edges of the GCC	2721
Diameter	15
Average Shortest Path Length	5.5215
Average Clustering Coefficient	0.0032

Table 2: Properties of Companies-to-Investors

Table 2 summarizes some of the basic metrics of our Company-Investor graph. Notice that most of the nodes fit in the Giant Component of this graph.

4.2.2 The Evaluation. For this project, we were interested in running our link prediction algorithm (Salsa) on the most important investors in the network. In other words, we wanted to see what are the companies that the biggest investors in our network should invest in (or should be suggested to invest in) based on the investments from similar investors. To find the most important nodes of type investor, we searched for the investors with the highest Betweenness Centrality, Degree Centrality, Closeness Centrality and Eigenvector Centrality. For all the metrics mentioned, 'Tencent Holdings' had the highest score. Also, 'Goldman Sachs', 'Sequoia Capital' and 'Temasek Holdings' were always in the top four investors with the highest score for each metric. So the four investors that were mentioned, were chosen to run the prediction algorithm on. We then run the Salsa algorithm we described in section 3.3 on each of the four investors to get a list of companies that should be suggested to them for further investments. To evaluate the performance of our link prediction, we wrote a script to find what percentage of investors in the circle of trust of the given investor are investing in the suggested company. This way, we could verify that the company nodes with the higher authority score were really the companies that received more investments from investors similar to the given investor. Here are the companies with highest authority score from our Salsa algorithm (Companies that should be suggested for further investments) for each of the selected investor:

- Tencent Holdings: 'Zipline' received the the highest authority score among the companies that they're not investing in. Also 42.85% of investors in its circle of trust are investing in 'Zipline'.
- Goldman Sachs: 'Roblox' received the the highest authority score among the companies that they're not investing in. Also 57.14% of investors in its circle of trust are investing in 'Roblox'.
- Sequoia Capital: 'Roblox' received the the highest authority score among the companies that they're not investing in. Also 57.14% of investors in its circle of trust are investing in 'Roblox'.
- Temasek Holdings: 'Nubank' received the the highest authority score among the companies that they're not investing in. Also 42.85% of investors in its circle of trust are investing in 'Nubank'.

We noticed that the companies mentioned above for further investments (ones with the highest authority score), received the highest score from our evaluation method among other companies we calculated the authority scores for. So that proves that our algorithm really suggests based on the behaviour of the investors similar to the given investor.

Evaluating Salsa Algorithm		
Investor	**Suggested Company**	**% of similar investors investing in the suggested company**
Tencent Holdings	Zipline	42.85
Goodman Sachs	Roblo	57.14
Sequoia Capital	Roblo	57.14
Temasek Holdings	Nubank	42.85

Table 3: Properties of Companies-to-Investors

Table 3 summarizes the evaluation results of our Salsa Algorithm.

5 CONCLUSIONS

For this project, the data set we found on Crunchbase had 630 unique unicorn startup companies, 985 unique investors and 18 unique companies(companies that are also investors). We read through previous work we found, analyzed their research process, method they used and results they concluded to understand the network better. We created a Companies-to-Investors undirected graph by using NetworkX. In this graph, nodes could have 3 types and edges were investments happening between companies and their investors. After we created the graph, we measured some of the basic graph metrics and visualize it in order to understand our graph and connections better. Based on our analysis (K-Means Clustering on node2vec features) we came up with a hypothesis for an investment recommender system that works based on a link prediction method called SALSA.

Also we can clearly see that in Software and Internet related Unicorn Startups, investors from America and China especially "Goldman Sachs" and "Tencent Holdings" have had the biggest influence in their surrounding companies, although there are also many more involved as well, while in Europe we see not much of influence happening in many industries and see that only in E-commerce "SoftBank Vision Fund" is the most influencial, and in some industries there are no influencers like Real Estate.

6 FUTURE WORK

- Creation of Investors-to-Investors graph where the nodes will only be investors. There will be an edge between these nodes if they funded the same company.
- Creation of Companies-to-Companies graph where the nodes will only be companies. There will be an edge between the nodes if two unicorn companies received funding from the same investor.
- Analyze the data from 2020 (Covid) and data from before 2020 (Pre-Covid), to see What impact has the pandemic had on the economy and investment patterns, we will check what industries' investment situation are affected or benefit because of the pandemic.
- Analyze the effect of the pandemic on companies and investors in different categories (i.e. Tech, Med, etc.)

REFERENCES

[1] 2020. *Crunchbase Private Unicorn Company List (ordered by recent funding).* Retrieved October 14, 2020 from https://www.crunchbase.com/lists/crunchbase-private-unicorn-company-list/f406c855-fff9-419b-9b44-324e1bfe3081/organization.companies?pageId=2_a_e22cde2f-eb8d-7dda-934c-557593f7bed7

[2] Adam Abdulhamid Charles Zhang, Ethan Chan. 2015. *Link Prediction in Bipartite Venture Capital Investment Network.* Retrieved October 8, 2020 from http://snap.stanford.edu/class/cs224w-2015/projects_2015/Link_Prediction_in_Bipartite_Venture_Capital_Investment_Networks.pdf

[3] James Chen. 2020. *Investor Definition.* Retrieved November 9, 2020 from https://www.investopedia.com/terms/i/investor.asp

[4] James Chen. 2020. *Unicorn.* Retrieved November 8, 2020 from https://www.investopedia.com/terms/u/unicorn.asp#:~:text=Unicorn%20is%20the%20term%20used,SpaceX%2C%20Robinhood%2C%20and%20SoFi.

[5] Manos Papagelis, Vanessa Murdock, and Roelof van Zwol. 2011. Individual Behavior and Social Influence in Online Social Systems. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia* (Eindhoven, The Netherlands) *(HT '11)*. Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/1995966.1995998

[6] Sam Schwagger. 2018. *Venture Capital Investment Networks: Creation and Analysis.* Retrieved October 9, 2020 from http://snap.stanford.edu/class/cs224w-2018/reports/CS224W-2018-38.pdf