COVID-19 and Unicorn Startups Investments Analysis - Milestone Report

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 Companies currently has gone through many changes. From our research on the effects of previous pandemics on the economy, we realized there is a profound demand of investments. Some scholars theorized this phenomenon as "likely wane, as labor scarcity in the economy suppresses the need for high investment"[8]. However, COVID-19 pandemic seemed to show opposing effects on the economy than previous pandemics. In other words, due to COVID-19, new investment opportunities were born, described by scholars as "opportunities for active investors"[7]. Hence, this motivated us to do an in-depth analysis of the effects of COVID-19 on the economy vs. previous pandemics, to predict new potential investments, and to finally reflect on this prediction in the real world setting.

In this project we used company-investor related data from Crunchbase [2]. 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 the years 2017 to 2020.

Ultimately, we believe that with more investments "the scope for future economic growth"[1] will increase. Therefore, we hope to influence the decision making process for new startup investment opportunities post-COVID-19.

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.

Having 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 [5].

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.[4]

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.

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.

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

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

2 RELATED WORK

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*[3]:

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, at least as of this milestone, we used the entire dataset for the prediction and evaluated our results through a automated process that we'll describe below. However, we plan to do more evaluations on our analysis in the final report.

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 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 our current milestone in the research (link prediction), 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 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 milestone.

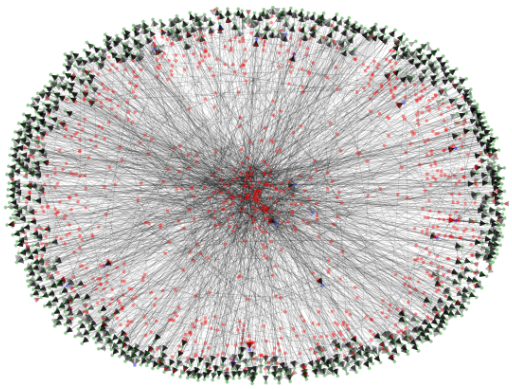

Figure 3.2.1: Visualization of the directed version of company-investor graph where the arrows go from investors to companies they invested in.

Figure 3.2.2: 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 3.2.1 visualizes the directed version of the Company-Investor graph. The arrows go from the investor to the company they invested in. Figure 3.2.2 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 3.2.2 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.

3.3 Link Prediction

The first major problem we wanted solve, was predicting links between companies and investors. 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

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

4.1 Basic Metrics

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

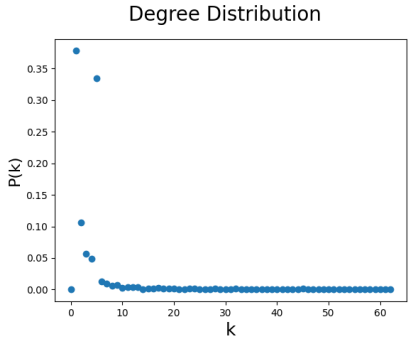

Figure 4.1.3: Degree Distribution of the Giant Component of the undirected graph.

Figure 4.1.3 shows the degree distribution of the graph. We can see that it is closer to a power law distribution.

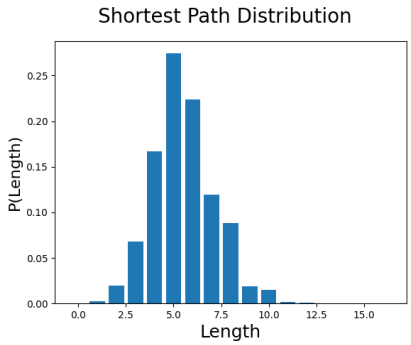

Figure 4.1.4: Shortest Path Distribution of the Giant Component of the undirected graph.

Figure 4.1.4 shows the shortest path distribution of the graph which has a binomial distribution. The average shortest path is 5.5 as noted in table 1.

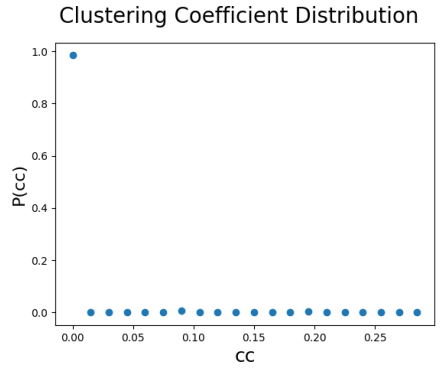

Figure 4.1.5: Clustering Coefficient Distribution of the Giant Component of the undirected graph.

Figure 4.1.5 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 1.

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 1: Properties of Companies-to-Investors

Table 1 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 Link Prediction 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 gtiven 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 2: Properties of Companies-to-Investors

Table 2 summarizes the evaluation results of our Salsa Algorithm.

5 CONCLUSIONS

For this milestone, 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). For this milestone, we read through previous work we found, analyzed their research process, method they used and results they concluded. Based on the understanding of their work, 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 got Degree Distribution, Shortest Path Distribution and Clustering Coefficient Distribution from it for future data analyze. We decided to do link prediction on what are the companies investors in our network may want to invest in. We used link prediction algorithm (Salsa) on the most important investors in our data, we searched for the investors with the highest Betweenness-Centrality, Degree-Centrality, Closeness-Centrality and Eigenvector-Centrality and found top companies with highest authority score from our Salsa algorithm to suggest to them for further investments. We evaluated them and they are all getting higher score than other companies(as Table2 shows), it proved the link prediction of our algorithm is reliable.

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

REFERENCES

[1] 2019. *Benefits of economic growth.* Retrieved October 7, 2020 from https://networkx.github.io/documentation/stable/reference/algorithms/clustering.html

[2] 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

[3] 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

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

[5] 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.

[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

[7] Franklin Templeton Investments Stephen Dover, CFA. 2020. *Active Investors Wake Up Before a Revolution.* Retrieved November 8, 2020 from https://advisoranalyst.com/2020/08/06/active-investors-wake-up-before-a-revolution.html

[8] Alan M. Taylor Òscar Jordà, Sanjay R. Singh. 2020. *Longer-Run Economic Consequences of Pandemics.* FEDERAL RESERVE BANK OF SAN FRANCISCO, SF, USA, Name of chapter: Theory: The natural rate and economic mechanisms after a pandemic, 1–4. https://doi.org/10.24148/wp2020-09