Startups and Their Unicorns

Ruth Johnson

March 15, 2016

CS 170A - Professor Parker

Contents

Introduction 1

1.1 Background

Entrepreneurship is growing, and more and more adults are starting to forego the traditional large company industries, and are flocking to the startup scene. in hopes of the next big startup. A few years ago, many of these entrepreneurs could only dream of their company reaching the 1 billion dollar mark. Today, these so called "unicorn" companies, or startups with a valuation greater or equal to 1 billion dollars, are rising. Companies like Snapchat and Uber grew to their current status in only a few years. Our goal is to analyze trends about these companies and see what sets them apart from the rest of the competition.

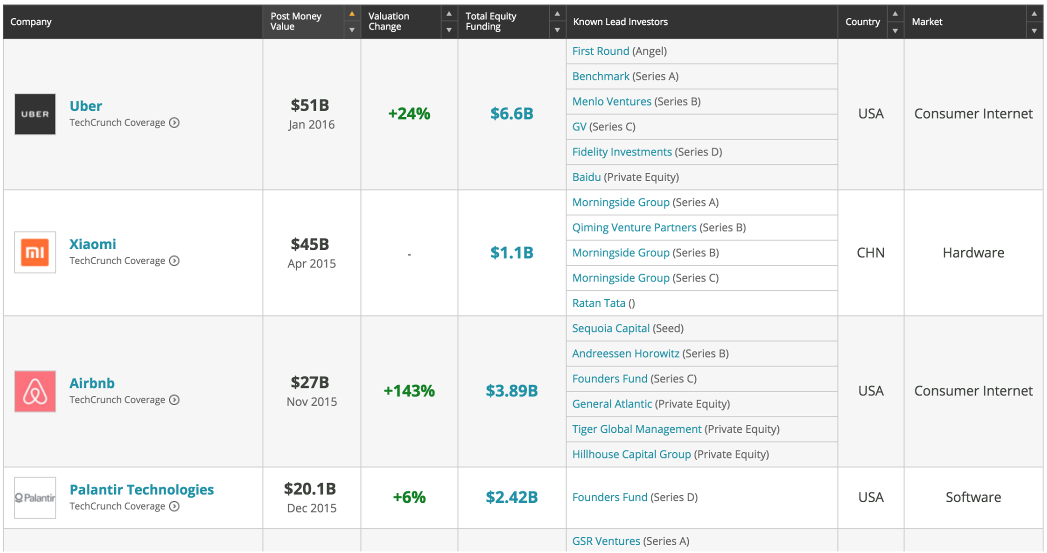
Unlike the dot-com bubble from the late 1990's and then in 2000, most of these small companies are not public, and are relatively young when compared to their larger, well-established competitors. A great deal of their money is in the private sector, coming from venture capitalists, accelerators, and other private sources. Because of this, data about startups are not all publicly available, but we were able to find significant trends from the data at hand described in the next section.

1.2 Data

The main dataset is from the 2015 Crunchbase Summary, which was retrieved using their API. Crunchbase is a crowd sourced database that is used to track startups covered on TechCrunch, a website that profiles startups and all startup news. We will use the current list of companies on Crunchbase and the information regarding their funding rounds, investors, and general information such as date of establishment and type of industry for our analysis.

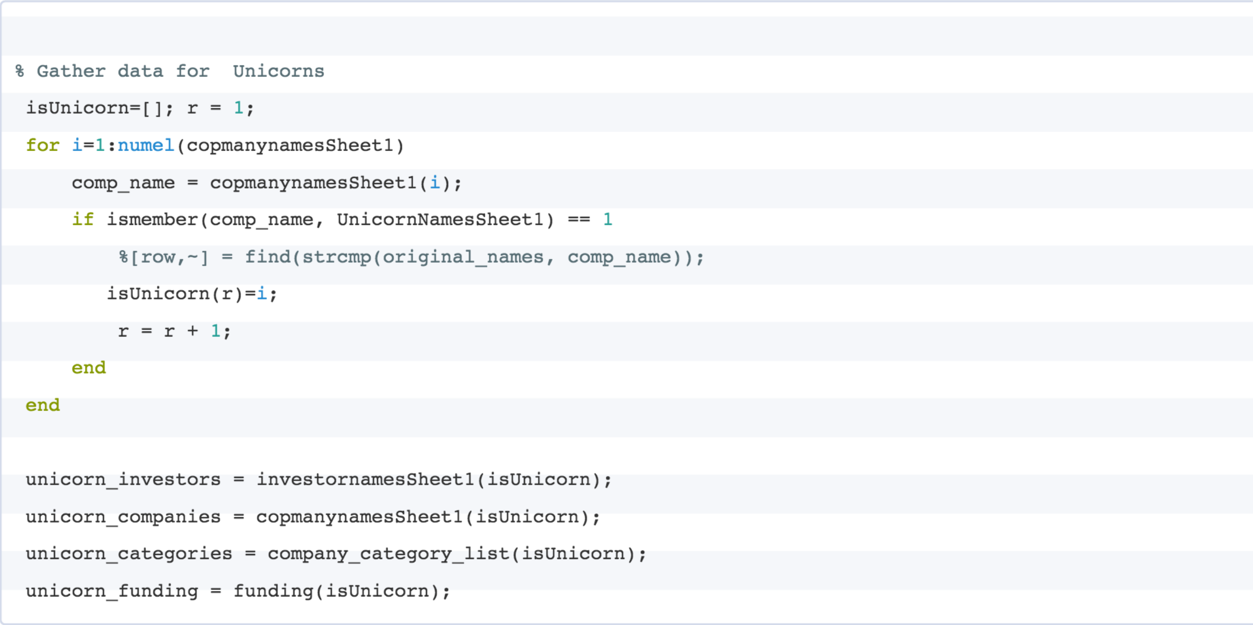
The *company* file contains the following categories: company name, company category list, country and/or state code, city, date founded, and total funding. The funding rounds dataset is comprised of the company name, each company's funding round types, funding round code, and amount raised in the specific round. Additionally, the investments file contains the information for each company's investors, such as date and amount invested. We plan to use all of these files as a comprehensive dataset to produce a larger picture of trends in the startup industry. Altogether, our dataset profiles approximately 50,000 companies from all around the world ranging from 1990 to 2015.

However, our current dataset did not have the list of valuations for each company. This figure is very important for estimating the success of a startup because by definition, a *unicorn startup* is a startup that has a valuation of $1 billion or higher. However, private companies (and more importantly their private investors) are not required to list their financial statements publicly, but some larger startups have estimated valuations that were gathered by TechCrunch researchers. TechCrunch released a list documenting the startups that have been inducted into the "Unicorn Club", and is updated hourly. However, this list was not available for download and lacked other valuable information that our larger dataset had. I was able to obtain the information on this list by web scraping the webpage pictured below, and then downloading it as a csv file.



Company	Post Money Value	Valuation Change	Total Equity Funding	Known Lead Investors	Country	Market
Uber TechCrunch Coverage	$51B Jan 2016	+24%	$6.6B	First Round (Angel) Benchmark (Series A) Menlo Ventures (Series B) GV (Series C) Fidelity Investments (Series D) Baidu (Private Equity)	USA	Consumer Internet
Xiaomi TechCrunch Coverage	$45B Apr 2015	-	$1.1B	Morningside Group (Series A) Qiming Venture Partners (Series B) Morningside Group (Series B) Morningside Group (Series C) Ratan Tata ()	CHN	Hardware
Airbnb TechCrunch Coverage	$27B Nov 2015	+143%	$3.89B	Sequoia Capital (Seed) Andreessen Horowitz (Series B) Founders Fund (Series C) General Atlantic (Private Equity) Tiger Global Management (Private Equity) Hillhouse Capital Group (Private Equity)	USA	Consumer Internet
Palantir Technologies TechCrunch Coverage	$20.1B Dec 2015	+6%	$2.42B	Founders Fund (Series D)	USA	Software
				GSR Ventures (Series A)		

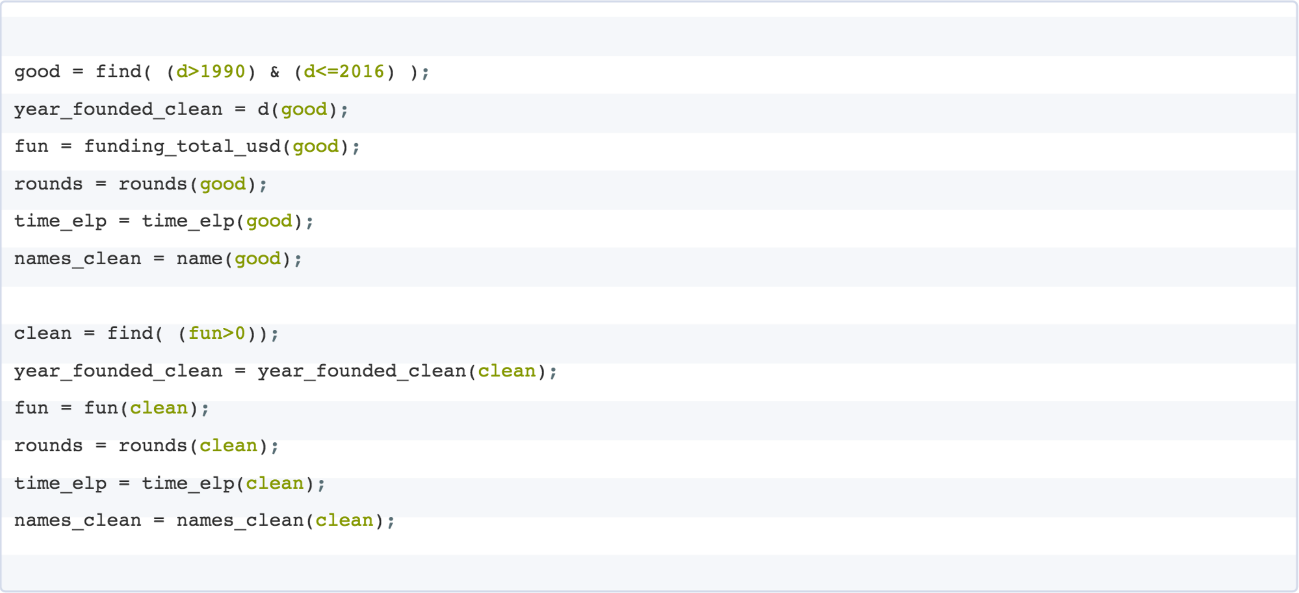
Then, by matching these names to the company names in our larger dataset, we were able to identify the corresponding indices of which companies were unicorn startups. Using this list of indices, we then parsed the other data categories ,such as investors, industries, and funding totals into a separate file, so that we have a dataset of just unicorn companies. Below is the code that shows how we were able to extract only the entries of the special unicorn companies from our larger dataset through the use of string comparisons.



```
% Gather data for  Unicorns
 isUnicorn=[]; r = 1;
 for i=1:numel(copmanynamesSheet1)
     comp_name = copmanynamesSheet1(i);
     if ismember(comp_name, UnicornNamesSheet1) == 1
         %[row,~] = find(strcmp(original_names, comp_name));
        isUnicorn(r)=i;
         r = r + 1;
     end
 end

 unicorn_investors = investornamesSheet1(isUnicorn);
 unicorn_companies = copmanynamesSheet1(isUnicorn);
 unicorn_categories = company_category_list(isUnicorn);
 unicorn_funding = funding(isUnicorn);
```

The data needed to be parsed to remove the empty entries and outliers. Some companies contained blank or invalid entries for certain categories; for example there were some blanks for *amount funded* because companies are not mandated to release this information. In addition, there were errors in the column listing for *date of establishment* because they were either after 2016 or came before 1990. When using this uncleaned data for analysis, we realized that this caused very visible outlines. By using Matlab and boolean expressions, the data was parsed to eliminate these entries in all corresponding categories. These errors could have possibly arose when the csv files were being created or when uploaded and delimited. After parsing, over 1,000 entries were removed, but we still had about 50,000 companies to use for our analysis.

The code below shows how we were able to quickly and efficiently parse all 50,000 entries.



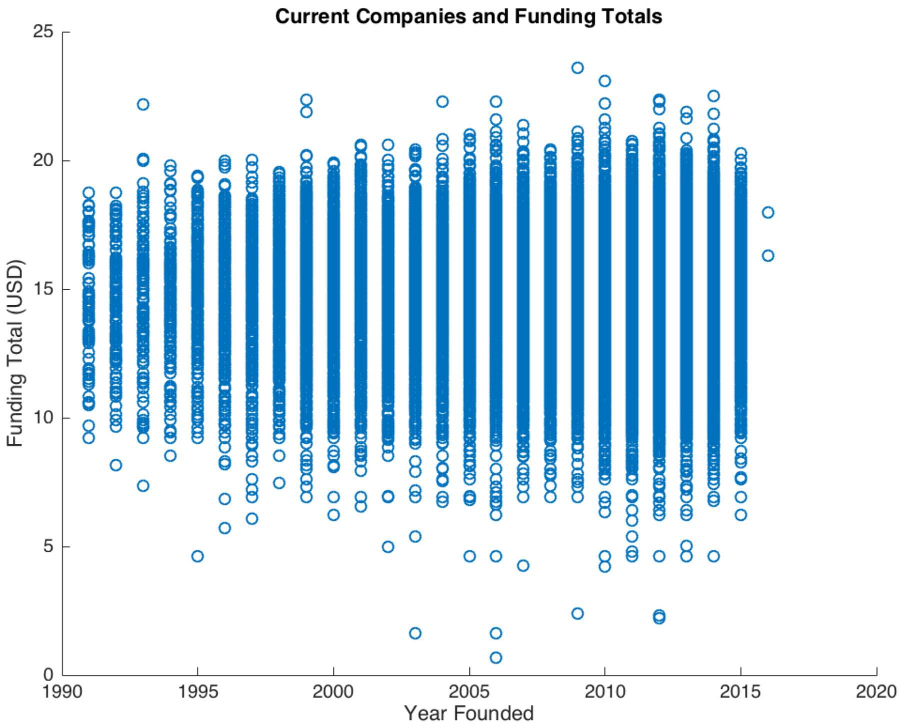
```
good = find( (d>1990) & (d<=2016) );
year_founded_clean = d(good);
fun = funding_total_usd(good);
rounds = rounds(good);
time_elp = time_elp(good);
names_clean = name(good);

clean = find( (fun>0));
year_founded_clean = year_founded_clean(clean);
fun = fun(clean);
rounds = rounds(clean);
time_elp = time_elp(clean);
names_clean = names_clean(clean);
```

By using this data, we hope to identify trends of very successful startups, and if they share any common characteristics. This is different than analyzing the trends of larger public companies because startups receive a great deal of their funding from private investors. Oftentimes, companies are measured by their financial statements, IPO scores, or stock prices, but almost all of these companies are not listed on NASDAQ or DOW, or have any other public record of these figures. Without these variables at hand, we had to come up with other parameters that were more accessible.

Analysis 2

2.1 Investment Trends

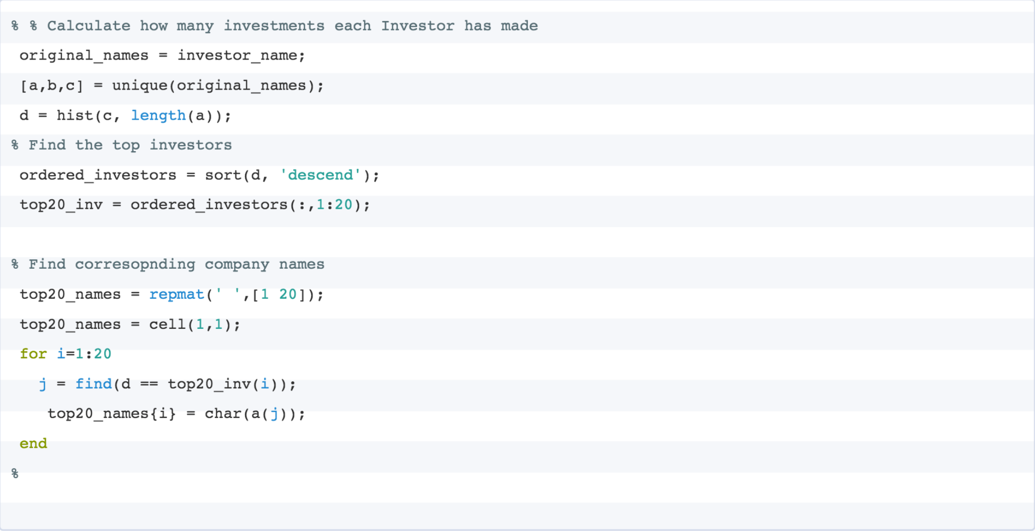
As mentioned previously, most of the startups get the majority of their funding from private investors. Companies have various rounds of funding; they propose their ideas, and investors can decide if they want to put their private money into the startup. This plays a crucial role in many startups because they need some source of money to start their company, and sometimes continued funds to keep it going until it can be sustained and maybe even go public.



Above are approximately all 50,000 companies plotted with their *year of establishment* versus their *total amount of funding*. Note though that the y-axis has been logged to modify the scale. One hypothesis would be that the older companies would have received more total funding since they have been around longer for more funding rounds from investors. However, the graph demonstrates that the younger companies, or roughly companies that have been established within the last

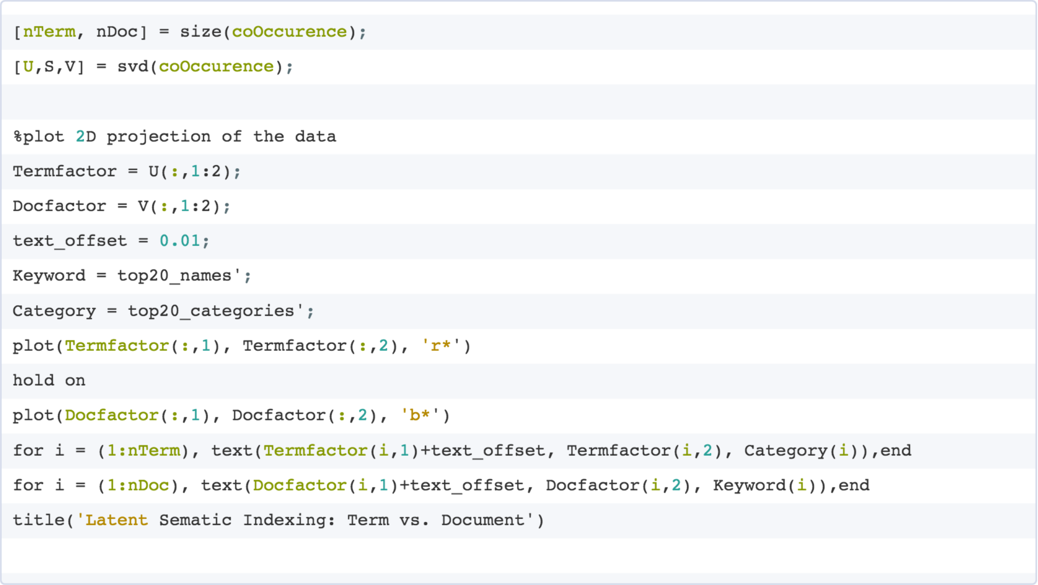
5 years have the most amount of funding. This means that there are many investors giving these early stage companies a great deal of funding. It appears that investors are willing to risk their money with companies that are not very established, and that in general, they are giving startups larger amounts of money than they have in the past. Could it just be that there are more unicorn statups now than 10 years ago because there's more private money being given out? We will see in a later section that this large amount of initial funding plays a large role in a company's success.

It seems that investors are more willing to invest their money into startups, but this still begs the question of where exactly their investments are going. We found the top 20 investors, the investors that invested in the most rounds, and the top 20 industries for our set of companies. In order to calculate the top investors and top industries, we had to calculate the occurrences of each investor and each industry. We did this by creating a unique vector that holds one instance of each investor, and then we parse the entire dataset of investments and counted how many times a certain investor had funded a company. Similarly, we were able to calculate the most popular industries using the same techniques. The code below demonstrates the scripts that were used.



```
% % Calculate how many investments each Investor has made
 original_names = investor_name;
 [a,b,c] = unique(original_names);
 d = hist(c, length(a));
% Find the top investors
 ordered_investors = sort(d, 'descend');
 top20_inv = ordered_investors(:,1:20);

% Find corresopnding company names
 top20_names = repmat(' ',[1 20]);
 top20_names = cell(1,1);
 for i=1:20
   j = find(d == top20_inv(i));
    top20_names{i} = char(a(j));
 end
%
```

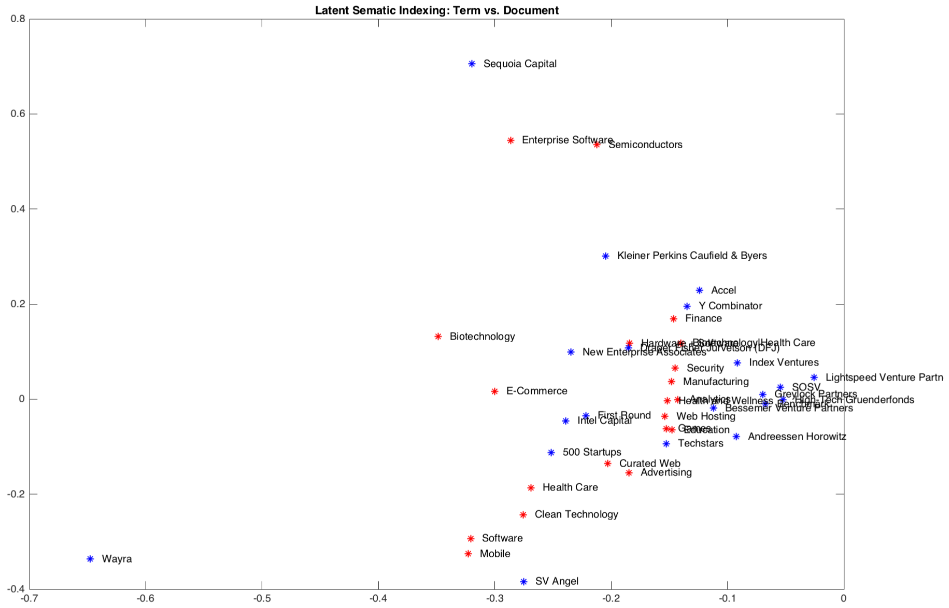
Using this data, we created a co-occurrence matrix, which showed for each of the top 20 investors, how many times they invested in one the top 20 industries. As shown below, using this frequency matrix, we created a 2D projection of the data.



```
[nTerm, nDoc] = size(coOccurence);
[U,S,V] = svd(coOccurence);

%plot 2D projection of the data
Termfactor = U(:,1:2);
Docfactor = V(:,1:2);
text_offset = 0.01;
Keyword = top20_names';
Category = top20_categories';
plot(Termfactor(:,1), Termfactor(:,2), 'r*')
hold on
plot(Docfactor(:,1), Docfactor(:,2), 'b*')
for i = (1:nTerm), text(Termfactor(i,1)+text_offset, Termfactor(i,2), Category(i)),end
for i = (1:nDoc), text(Docfactor(i,1)+text_offset, Docfactor(i,2), Keyword(i)),end
title('Latent Sematic Indexing: Term vs. Document')
```

By graphing the projection, we can see the relationship between each of the top 20 investors and the industries near it on the graph. A small distance between an investor and an industry on the graph means that these are the industries that the investor tends towards to invest the most in, when compared to the other industries.

A significant investor is Sequoia Capital, who had the most investments made out of all of the investors. We can see that they tend to put a great deal of their money into Enterprise Software and Semiconductors. Notable unicorns invested in by Sequoia Capital include Github, Evernote, and Docker, as well as various others. Another interesting point is on the far left, Wayra, an investor that does not seem to make very many investments in the top 20 industries. However, there is a clump of both industries and investors on the right side, suggesting that these investors are all equally investing in these popular industries.

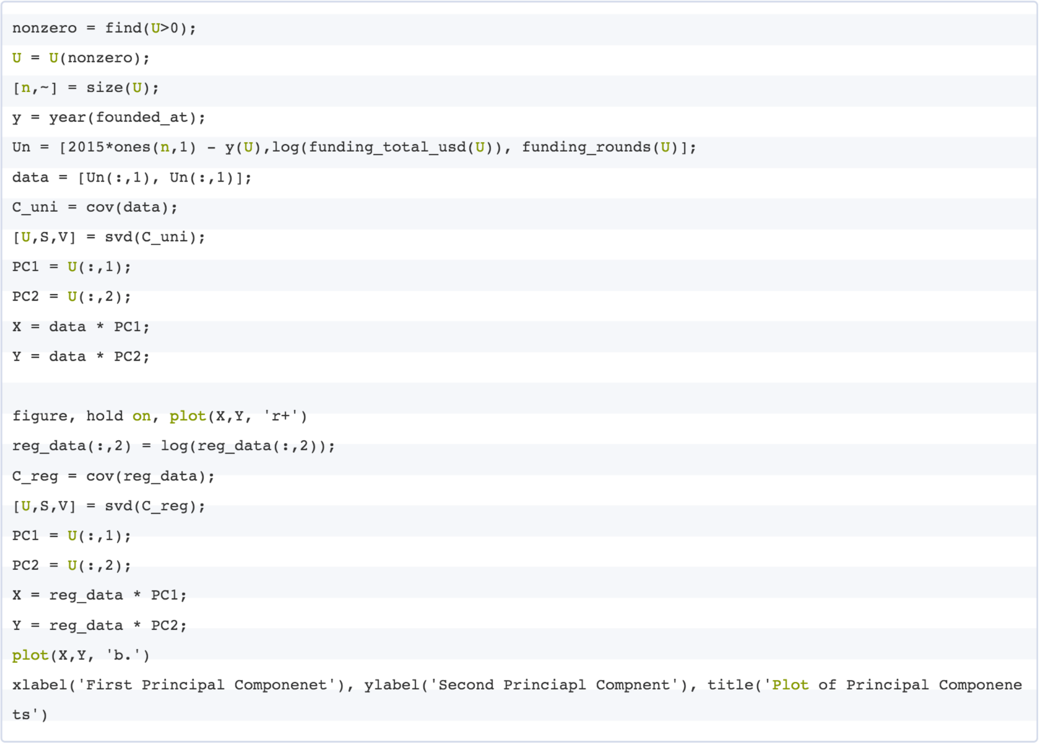


From our analysis thus far, we can see a trend of investors recently giving large investments to very young companies, as well as what types of companies. The top 5 industries were: biotechnology, software, mobile, enterprise software, and E-commerce. This gives us an idea of the current trends for the startup market, which is useful for identifying areas of possible future success.

2.2 The Rich Get Richer

Next, we want to identify any possible trends of these very successful startups. By doing a principal component analysis, we can see which variables give us the best overall approximation of our dataset. We first found the covariance matrix of our data, and then took the SVD to find the first and second principal components. Then, graphing it as a scatterplot shows us a visualization of any trends in our companies. In the graph below, the first principal component emphasizes years established, so this has the highest variance among all linear combinations. The second principal component emphasizes the total amount of funding. Recall, that the funding totals were logged in order to scale it appropriately when comparing it to other values.

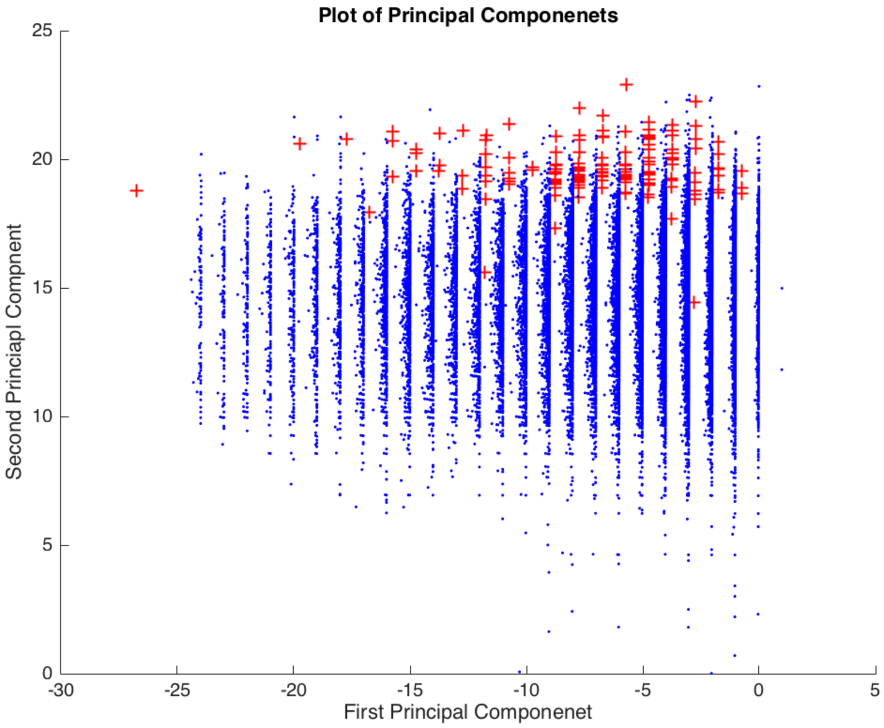
The code below shows how this was calculated using MATLAB.



```
nonzero = find(U>0);
U = U(nonzero);
[n,~] = size(U);
y = year(founded_at);
Un = [2015*ones(n,1) - y(U),log(funding_total_usd(U)), funding_rounds(U)];
data = [Un(:,1), Un(:,1)];
C_uni = cov(data);
[U,S,V] = svd(C_uni);
PC1 = U(:,1);
PC2 = U(:,2);
X = data * PC1;
Y = data * PC2;

figure, hold on, plot(X,Y, 'r+')
reg_data(:,2) = log(reg_data(:,2));
C_reg = cov(reg_data);
[U,S,V] = svd(C_reg);
PC1 = U(:,1);
PC2 = U(:,2);
X = reg_data * PC1;
Y = reg_data * PC2;
plot(X,Y, 'b.')
xlabel('First Principal Componenet'), ylabel('Second Principapl Compnent'), title('Plot of Principal Componene
ts')
```

The Unicorn startups have been plotted with a red ’x’, and the other companies are in blue dots in the graph below. Clearly, we see the Unicorn startups forming a cluster at the top right portion of the graph. Additionally, there is little spread of the markers in red on the x-axis when compared to the blue.

It is important to note these startups in the red, or the Unicorns, are in a separate group due to their valuation, not their total funding. This means that these companies show that they have both a very high valuation as well as have received a large amount of funding already. This follows the "rich get richer" trend, or that the companies that started with a large amount of money are making the most money, which then gives them the resources to make even more. As our previous analysis had shown, the newer companies are receiving more funding, which could help them to a higher valuation in the future.



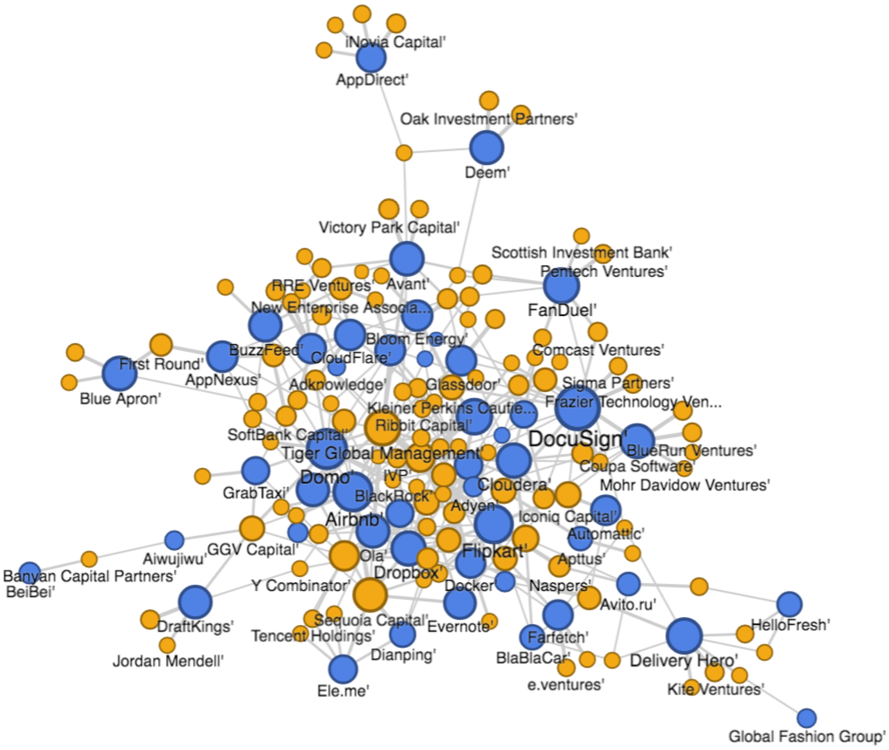
However, a company with a high funding total does not necessarily have a high valuation, since there are still dots in the blue that are intermixed with the red x's, denoting that these companies have the save or even more funding as the unicorn companies, but their valuation has not yet reached $1 billion. This means that there are still factors that will either help or hurt a company from reaching that $1 billion valuation mark other than the amount of funding they receive.

2.3 It's All About Who You Know

After analyzing the most popular industries, investors, and the amounts of funding, there still seems to be a missing connection that differentiates the unicorn startups from the rest of the startups. A commonality between almost all of the unicorn startups were their high amount of funding, but we also saw that this large amount of funding did not guarantee a high valuation. Next, we turn back to the source of this funding, the investors again.

Using Google's Fusion Tables, we can visualize all of the investments for the unicorn startups. We made a table of all of the investments made to a unicorn startup. Then, the Google Fusion Table API placed each company and investor as a node; the graph below shows each of the Unicorn startups in blue, and the investors in yellow. A connection on the graph is made between the company that an investor has funded.

The more connections each node has, the greater the radius of the node displayed. The graph itself is quite dense; this shows that there are a lot of connections between the investors, or that many of the investors are funding multiple unicorns.

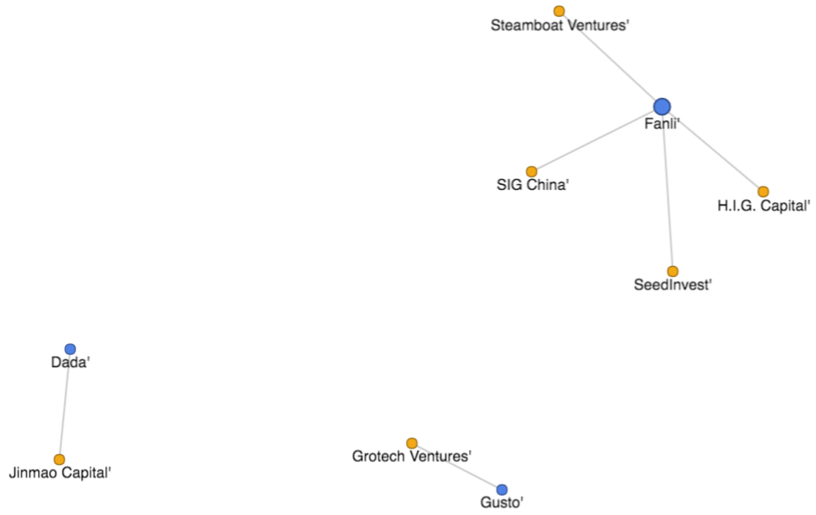


The interactive version of the graph can be found here or accessed at the link below, where the nodes can be moved and zoomed in or out, which allows for a clearer picture of the data.

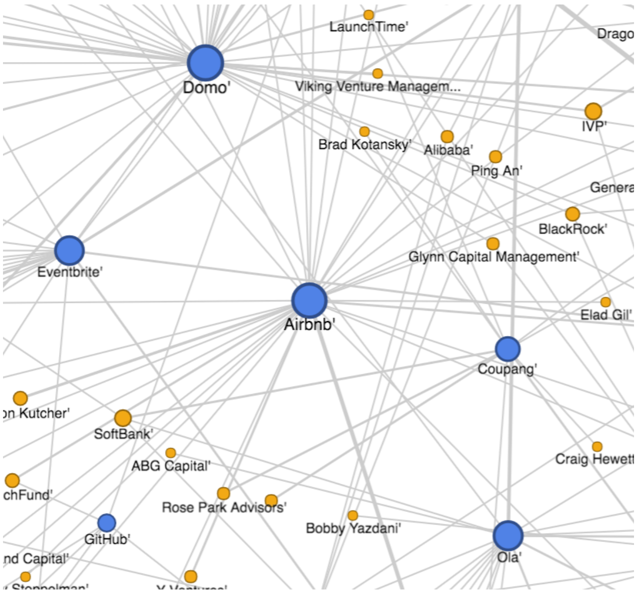
$https://www.google.com/fusiontables/DataSource?docid = 1uq93nzr5E06qjGw$
$AHGpU_wHj6hPlRPb_TxyLX3Mw\#chartnew : id = 3$

We see a remarkable network made of almost all of the unicorns. This graph indicates that something of similarity between all of the unicorns are the investors backing them up. It seems that there is a subset of investors that are the primary investors for unicorns. The numerous connections in our graph are very significant because this means that they are funding for the most part, more than one unicorn. We have found another common characteristic among our data.

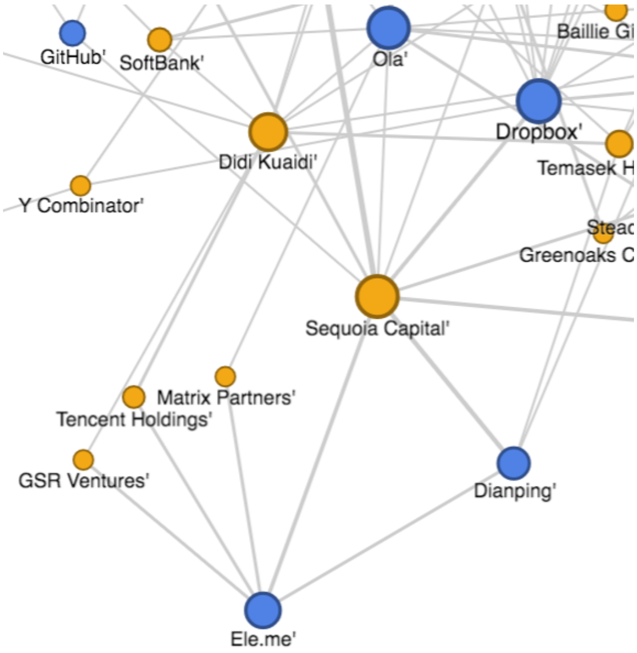
However, there are some investors that are only funding one unicorn, however, out of the 159 official unicorns reported, only 3 companies are disconnected from the graph (not pictured in the main cluster above). We see that for the most part, almost all of the unicorn startups are connected by their investors, and this means that approximately 98% of the companies share at least one investor with another unicorn startup. Below, we can see pictured, the 3 companies: Dada, Gusto, and Fanil'.



Taking a closer look at one of the companies, Airbnb in this case, we see how many connections it has. Not only does this demonstrate the Airbnb has a lot of investors, but this also shows that its investors are also investing in other successful startups. In the larger graph, it is located in the dense clump near the bottom left-hand side.



Additionally, zooming in on the investors, in this case Sequoia Capital, we see that it has been investing in a total of 9 unicorn startups. This means that they are currently backing 5.66% of the total unicorn startups. This may seem like a small percentage, but this also means this small grouping of their companies are valued over $9 billion dollars.



These graphs show that the majority of unicorn startups tend to be clustered around many of the same investors. This could mean a variety of things; one possible hypothesis is that this subset of investors know how to work with these early stage startups and exponentially increase their value since they have had experience doing so in the past. It is definitely a challenge to break the $1 billion valuation mark, but it may be easier if your investors already have worked with companies that are already in this exclusive clustering. The alternative hypothesis is also that these investors will not invest unless they feel like the company has a true potential to be a success story. Either way though, it is very interested that this clustering has occurred. This supports our "richer getting richer" hypothesis form the previous section, except this time with the investors. It seems that those investors who end up funding the most successful companies are those who have a history doing so. Additionally, this supports the idea that is not solely a company's numbers that can lead them to success.

Conclusion 3

3.1 Summary

Our goal was to try and find common trends of the unicorn startups when compared to other startups. A large challenge was trying to find adequate enough public data to use because private companies are not required to release their financial statements, and many smaller startups do not have very much data yet, at least not publicly. By analyzing our given data and combining other sources, such as the list of unicorn startups we had webscraped, we were able to come up with a dataset that could produce clear results.

We first wanted to get an idea of what the startup market currently looks like. After graphing the amount of funding that companies are receiving presently when compared to 10 years ago, we can see that there is a steady rise in the amount of funding these companies are receiving. Since startups are almost financially supported through these private investors, we can see that there is a lot of money in the private sector now and for the last few years. In addition, by projecting the top investors and top industries, we could see in what industries the market is leaning towards, as well as who exactly is supporting which industry. Many of the unicorn companies fell within the top 20 industries and had some of the top 20 investors.

Through our analysis we found a trend of many of these unicorn startups also having some of the highest amount of funding. This suggests the pattern that those companies that receive a lot of money from their private investors will continue to make more money. Additionally, many of the younger startups are the ones receiving this funding, which is where a lot of the startup unicorns fall, being established only 5 or less years ago. Although we cannot track the valuations of all companies, we were able to see that using the amount of funding received thus far is a strong indicator of a high valuation.

Lastly, we found that many of these unicorn startups share many of the same investors, with only 3 companies not sharing a single investor. This shows strong evidence that the investor has a lot to do with the success of a company. Not only do these investors provide the necessary funds, but they also provide connections to the startups. It could be that these investors work with the company to help them achieve such a high valuation, or possibly that they will only choose to invest in companies that they forsee will be very successful. Either way, our data shows that the unicorns all show certain trends with their investments, through the amount invested and even who is investing.

Although this information cannot be used to predict which companies will be next to join the ranks of the $1 billion valued companies, like the unicorns, it demonstrates trends among the current successful startups, which could be useful in predicting successful qualities for a potential company.

3.2 Future Analysis

There are many extensions of this project that can be made in the future. One limitation that was mentioned previously was accessibility to public company data. There were resources that could provide more in-depth data, but these resources required purchase or special permissions.

An additional extension would be to see if social media has a significant impact on a company's success. On CBS Insights, a website that also profiles startups and venture capitalists, they track companies online presence through their social media, mentions in news articles, adds/campaigns, etc. I think this would be a very interesting analysis since many of these startups grow so quickly, so one would predict that it might have to do with their digital presence that makes them so well-known so quickly.