Features of Startup Founders – Exploratory Analysiss

Mei Fu

4/16/2017

Contents

0.1 Motivation

Innovation is always one of the key momentums of social development, and startups take a leading role in the innovation. The success of a startup is largely related to its founders especially during the early stage, but it is relatively hard to know the features of successful founders through traditional way.

This project is interested in studying some characters of founders who have a higher probability of success or more favored by venture capitals. To achieve this goal, I searched several available investment databases including CrunchBase, Bloomberg, FactSet, etc. Considering the limitation of account access and the nature of the dataset, I chose to crawl data from Angellist.com (which is ok after checking /robots.txt), Linkedin, and Twitter mainly using Python and Selenium package.

The data acquired belongs to three aspects: social networks, skills, and influence power in order to illustrate the founders' characters, which could be useful for those who have entrepreneurial ideas or angel investors. Specifically, this project is going to study the following questions:

1) Do founders likely to have a similar background? How do their social networks look like?
2) What are the most frequent schools or companies?
3) What are the most important or popular skills to be a good founder? What are the founders good at?
4) How is their influence power on the social media? Do they have a lot of tweets, followers, retweets, likes?

0.2 Data Source & Collection

The basic logic of the data gathering part is as follows (Figure 1):

1) write **StartupsURL.py** to filter data down to a reasonable size, get startups basic information including name, page URLs, etc., and save them into the first CSV file **startups_info.csv**;
2) write **FoundersnFundings.py** to go to the startup page recorded from step 1, and save the funding details into a CSV file **startups_funding.csv**;

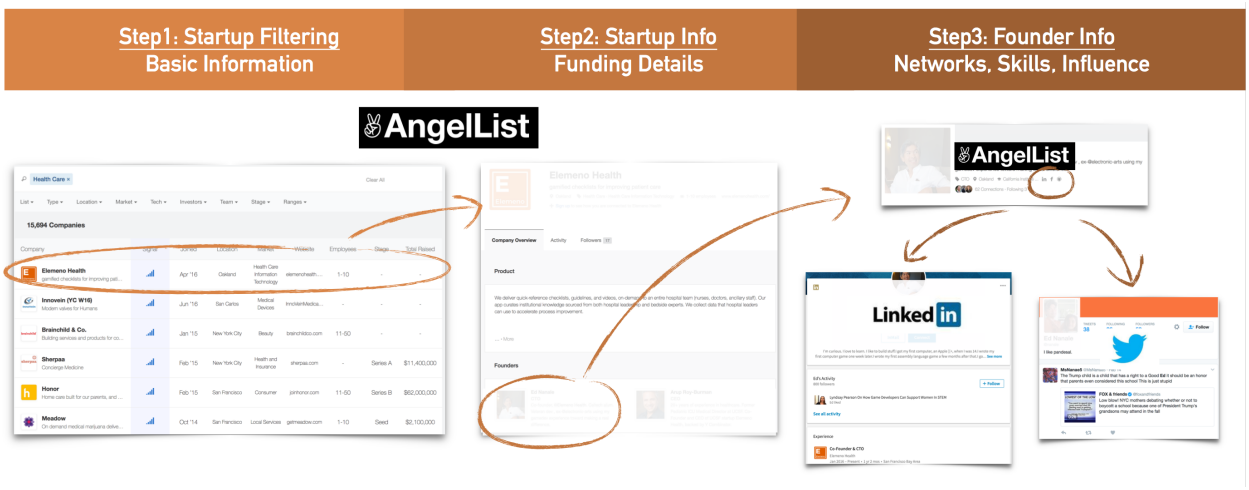
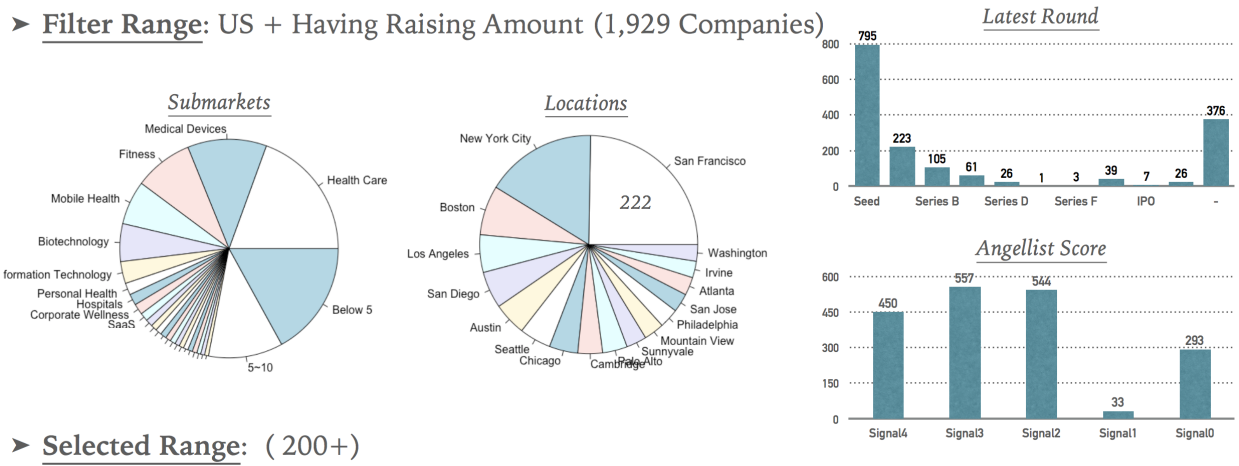


Figure 1: The Flow of Data Collection

➤ **Industry**: Health Care (15,693 Companies)

➤ **Filter Range**: US + Having Raising Amount (1,929 Companies)



➤ **Selected Range**: (200+)

Figure 2: A Glance at the Dataset

3) further use **FoundersnFundings.py** to get founder's linkedin & twitter URLs on the page, go to them, and save the schools, companies, skills, and tweets into **startups_founder_skills.csv**, **startups_founder_networks.csv**, and **startups_founder_influence.csv**.

4) use **main.py** to run the code.

Due to the huge available data size and the waiting period between each call, the data were focused in the US health care industry, and the whole process was divided into 6 parts to run parallelly using different machines. There were about 1,900 startups saved, within which around 1,700 funding information and 250 founders' information were retrieved and stored in the CSV files. (Figure 2)

0.2.1 Angellist (https://angel.co)

The Angellist.com is a well-known website with relatively detailed information on the early stage startups, founders, and investors. The information on the website is generally for free. Since it stops accepting new API applications, I chose to crawl the data directly from the website with a reasonable time interval between each call (20 ~ 30 seconds).

Firstly I imported the Selenium package, which is a portable software testing framework for web application. I set up a Selenium browser by defining **set_browser(size1, size2, secs)**. The inputs are the width and height of the browser and the potential waiting time for a web page. The function returns a PhantomJS web driver object and the waiting time. Then I defined **get_full_url(market)** to get 6 URLs for the company filtering pages based on the different ranges of the raising amount, which allowed me to acquire more detailed data parallelly. The input is the industry of the startups (such as 'Health Care' in this project), and it returns a list of URLs for the filtering pages according to the predefined selecting criteria (such as different ranges of raising amount in this project). The key point of setting filtering criteria here is to make sure each page will return balanced amount of startup information and not exceed the searching limitation at once, which is about 400 companies for Angellist. Using these URLs, I then created **get_startups_url(market, url_num)** to save the basic startups information including their Angellist page URLs.

The 6 filtering URLs are as follows (the later range excludes the former range):
url1: raising amount < $100k
url2: raising amount < $500k
url3: raising amount < $2m
url4: raising amount < $10m
url5: raising amount < $80m
url6: raising amount < $100b

The URLs for the startups' web pages are like this:
https://angel.co/breakthrough-com

The URLs for the founders' web pages are like this:
https://angel.co/goldenson

The data returned from the web pages were all in the string format, and dictionaries were created to store those strings. The most important variable (key) for this step is the URL for the startup pages, because it is used to get further detailed information in step 2 and 3. (Figure 3) I defined **load_startup_urls(url_num)** to open the saved CSV file and get a list of startup URLs. The URLs were sorted so that it would be possible to resume from the latest record when losing the Internet or blocked by the website.

For step two, I defined **get_founder_funding(url_num)** to retrieve the data. The idea was to use the browser defined in step one to open the startup URLs, and to find the related HTML tags by using Selenium methods similar to the Beautiful Soup. The most important funding variable (key) is the funding round so that the tweets information can be compared according to the different funding rounds assuming that the later the funding round is, the more successful the startup is. The other variables such as news, date, type, investors are all worth further mining. But these features are more related to the startup analysis, so they are not the focus of this project, but I kept them in the dataset in case of future usage. (Figure 4)

Company Name	URL	Pitch	Signal	Location	Markets	Website	Employees	Stage	Total Raised
Genomera	https://angel.co/genom	Health Studies & Clinical Trials at	Signal4	Silicon Valley	Clinical Trials	http://genomera.com/		-	$20000
CUR	https://angel.co/cur	A smart band-aid for pain relief	Signal4	San Francisco	Medical Devic	http://www.cur.me/	1-10		$1
Reify Health	https://angel.co/reify-h	Building a more creative healthcar	Signal4	Boston	Medical Devic	http://www.reifyhealt	1-10		$20000
Smart Patients	https://angel.co/smart-	communities for smart cancer pati	Signal4	Mountain View	Personal Hea	http://www.smartpati	1-10	Seed	$100000
Quantified Care	https://angel.co/quanti	Better health starts at home	Signal4	Baltimore	Mobile Health	http://www.quantifiec	1-10		$50000
Paubox	https://angel.co/paubo	The easiest way to send and recei	Signal4	San Francisco	Enterprise Se	https://www.paubox.	1-10	Seed	$30000
Lifey	https://angel.co/lifeyba	Providing faster emergency respo	Signal4	San Francisco	Lifestyle Busi	http://www.lifeyband	1-10		$10000

Figure 3: Glance at startups_info.csv

Company Name	Round	Type	Date	Amount	News	Investors	Investor Location
1776	0	No Stage	Apr 1 2016	7200000	http://www.1776.vc/press	K Street CapitallKidc	WashingtonlWashing
(GlobalMusic4 LIfe) (	0	No Stage	Jan 16 2015	700	https://fundly.com/globalr	Joanne Chan	
100Plus	1	Seed	Nov 2 2011	500000	http://www.finsmes.com/2011/11/100plus-receives-initial-funding.htm		
100Plus	0	Seed	Nov 30 2011	750000	http://techcrunch.com/20	Greylock PartnerslB	Menlo ParklMenlo P
1DayMakeover	0	No Stage	Jun 30 2008	50000			

Figure 4: Glance at startups_funding.csv

For step three, the Founders' LinkedIn and Tweeter URLs were called directedly when running the program, so these URLs were not included in the output of this step. Although Angellist also includes founders' basic information including schools and former experience, I assume the information on the Linkedin would be more up-to-date.

0.2.2 Linkedin (https://www.linkedin.com)

Linkedin doesn't provide free API that meets my need, so the data were also crawled with a time interval between 30 ~ 40 seconds. The LinkedIn data included founders' social networks and skills features, which were also in string format and were stored into dictionaries. I used Beautiful Soup to find tags here because it is more suitable to crawl the HTML page. The most important variables in this part are schools (since college) and former companies to show social networks, and self-defined skills, majors, and former positions. These information could show founders' social networks and skill-sets. The data were saved into CSV files.

0.2.3 Twitter (https://twitter.com)

Although Twitter has powerful APIs like Tweepy, using Selenium is more consistent with the whole workflow, so I still used it and used CSS class and XPath to locate the tags here. The important variables of this part are founders' followers, tweets, likes, and retweets numbers. These variables can show founders' influence power on the social media.

After all these steps were done, the parallelly saved CSV files were combined into one CSV file by using R rbind method.

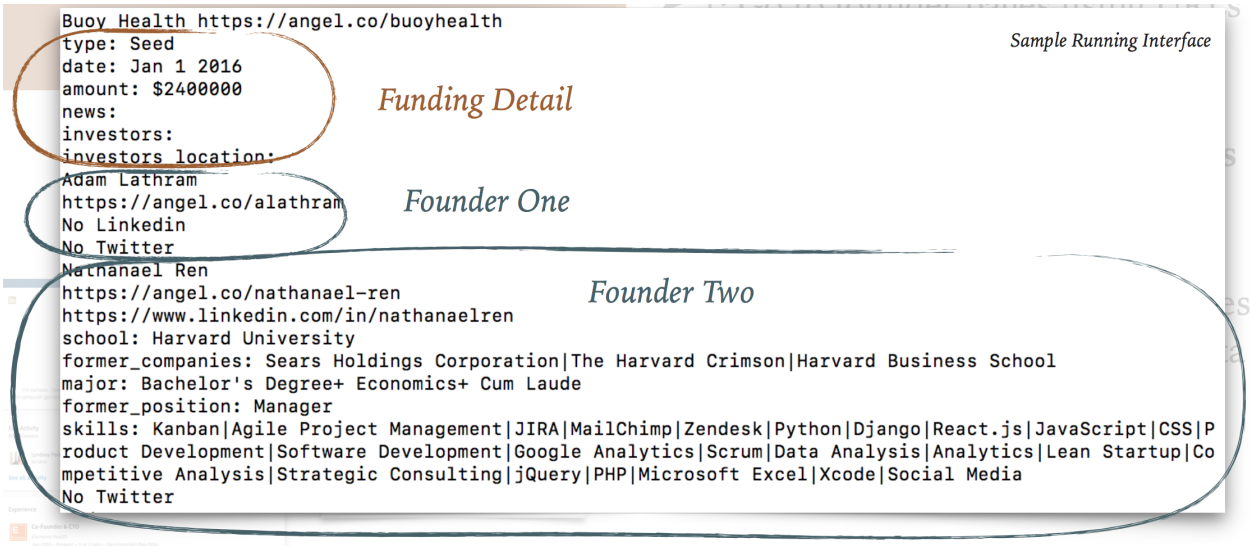


Figure 5: A Sample Interface when Running the Program

0.3 Data Manipulation Methods

0.3.1 Data Cleaning

The first step of data manipulation was to clean the data. Since I retrieved the data directly from the webpage, I handled the missing value issue within the collection phase by using 'try-except'. If the startup doesn't have funding records or the founders don't have Linkedin or Twitter URLs, those startups or founders won't be recorded in the saved CSV files. If they did have these URLs but the web pages were private or couldn't be visited by a spider, I would skip these founders and startups as well. Actually, there were a lot of founders without Twitter links or Linkedin URLs on Angellist. Skipping these saved me a lot of time since this is a pretty time-consuming process. The missing values in the saved data such as news or investors were not important to the main goal, so I didn't conduct extra steps to deal with them and just left them blank in the CSV files.

But some of the format issues do need to deal with such as '63.2K' should be converted to '63200' when recording tweet counts, and also changing the raw college names under the same rules. These cleaning steps were conducted directly in the CSV files. On one hand, it was more time saving to modify minor or uncommon issues directly. On the other hand, the raw college names were inputted by founders themselves and sometimes ambiguous, so it was difficult to set arbitrary rules to modify them by using programs. However, if the data size scaled and this step needs to be repeated, it would be more efficient to spend time figuring out common modification rules or calculating edit distance or similarity in Python. (As what I did in part B)

0.3.2 Data Combination

The second step of data manipulation was to combine and manage the datasets so that they could be easily accessible. I also handled a part of the merging tasks within the collection process, since I integrated the collection process from Angellist startup pages, their founders' Linkedin pages and Twitter pages within one function call.

So the rest of the combination process was to store all the CSV files into a database. I mainly used Sqlite3 to create a local database **startups.db**. The data from each saved CSV files were inserted into the 6 tables accordingly in the database, and the data can be joined across the tables based on the company_name as the relational unique key. In order to do the visualization in part A, I selected needed data (schools, startups, funding rounds) by joining the networks table and the funding table using SQL queries and stored them into dictionaries to do visualization and analysis.

0.3.3 Data Processing for 4 Questions

Thanks for all the data manipulation and combination process completed in part A, I was able to load the CSV files directly into R and do the data analysis.

Specifically, for question 1, I used iGraph library to draw 3 graphs to explore the founders' networks. Firstly, I loaded 'startups_founder_networks_total_companies.csv' and cleaned data within a for loop to separate the school data by using str_split(data, '\|') because the original data were stored using the mark '|' to join different attended schools or companies for each founder within a single string. Then I stored the tuple of the startup, founder, school into a list and ultimately transformed into a data frame to record the relationship among those variables. I then built 2 node sets — founders & schools — and added their relationship data as the edges.

After that, I used iGraph to draw the bipartite graph of the founders on one side, and the schools they attended on the other side. Finally, I drew a projection graph of the founders based on the connection of the bipartite graph and used Walktrap method to illustrate communities within the network. The idea is basically that given a starting node, what ranges will the node reach within n random walk steps. The R shiny was used to draw an interactive graph with different random walk steps for Walktrap. Similarly, I also used this method to process and draw the founders' network graph and the comprehensive network graph.

For question 2, basically, I used data from question 1 to count and show the most frequent schools and companies attended by the founders. After filtering the top former companies, I manually entered their industry labels to add another important dimension of the features and making the graph more informative.

For questions 3, I used the data in 'startups_founder_skills_total.csv'. In order to know the most important or popular skills to be a good founder, I further separated the question into 4 sub features – their degrees, undergraduate or graduate majors, self-defined skills, and their latest former positions. The degrees were not easy to deal with. The self-report data, especially the text data, always facing the problem of ambiguity in semantics and diversification in morphology. Unlike using the same rule to manually modify the school names in part A, I predefined the data related to 'MBA', 'doctorial', 'ba', 'master', etc., and used the 'sub' function to combine them together.

For the last part, I used 'startups_founder_influence_funding.csv' dataset. This dataset is clean, so the manipulation in this part is mainly to omit the NA values before plotting.

Most of the difficulties in this part were related to handling diverse text data in question 3, and I solved this by substitute certain predefined words. Besides that, most of the learning efforts were to explore other R libraries, such as R shiny and iGraph, which were not covered in class but would be useful for this project. IGraph is very powerful in analyzing network graphs, and R Shiny is good at plotting interactive graphs, which is helpful to show how the communities would like when using a

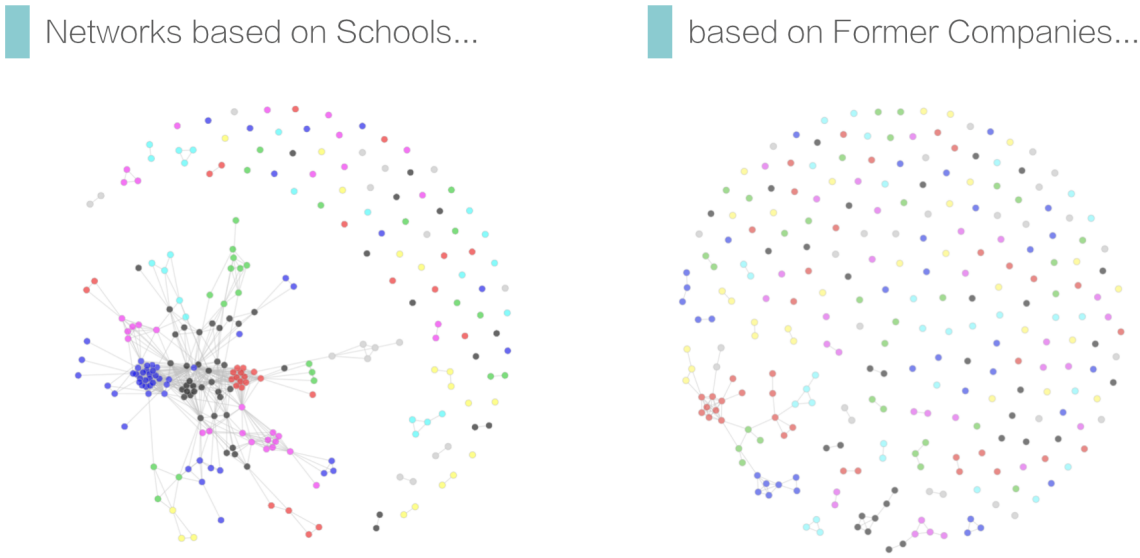


Figure 6: Founder networks based on Schools and Former Companies

different parameter. Reading through the documentation and using StackOverflow to find tips were the most useful resources when learning a new library or getting stuck into a problem in the project.

0.4 Analysis and Results

0.4.1 Question 1: Do founders likely to have a similar background? How do their social networks look like?

I drew several network graphs to analyze the different relationships of founders according to graduation schools, former companies, and current startup colleague. (Figure 6) We can see that obviously the founders tend to have a close relationship based on their schools, which is intuitive. 60% founders are connected in the largest group, which indicates that they are likely to go to great schools, and great schools are not many. (will dig into this point in question 2) There are several sub-communities within the group because some founders who went to different schools act as the bridges between the communities.

The founder network based on their former companies was surprising to me at first glance since it is pretty sparse. Only 14% founders are connected in the largest group. However, it also makes sense because the choice of companies are much wider in terms of different industries, and the founders also tend to have former startup experiences.

After considering all these relationships together, the connection within the largest component increased to 82%, and on average they can reach each other through around 3 people. Two major communities formed within the group, which indicates that their connections are more close. (Figure 7)

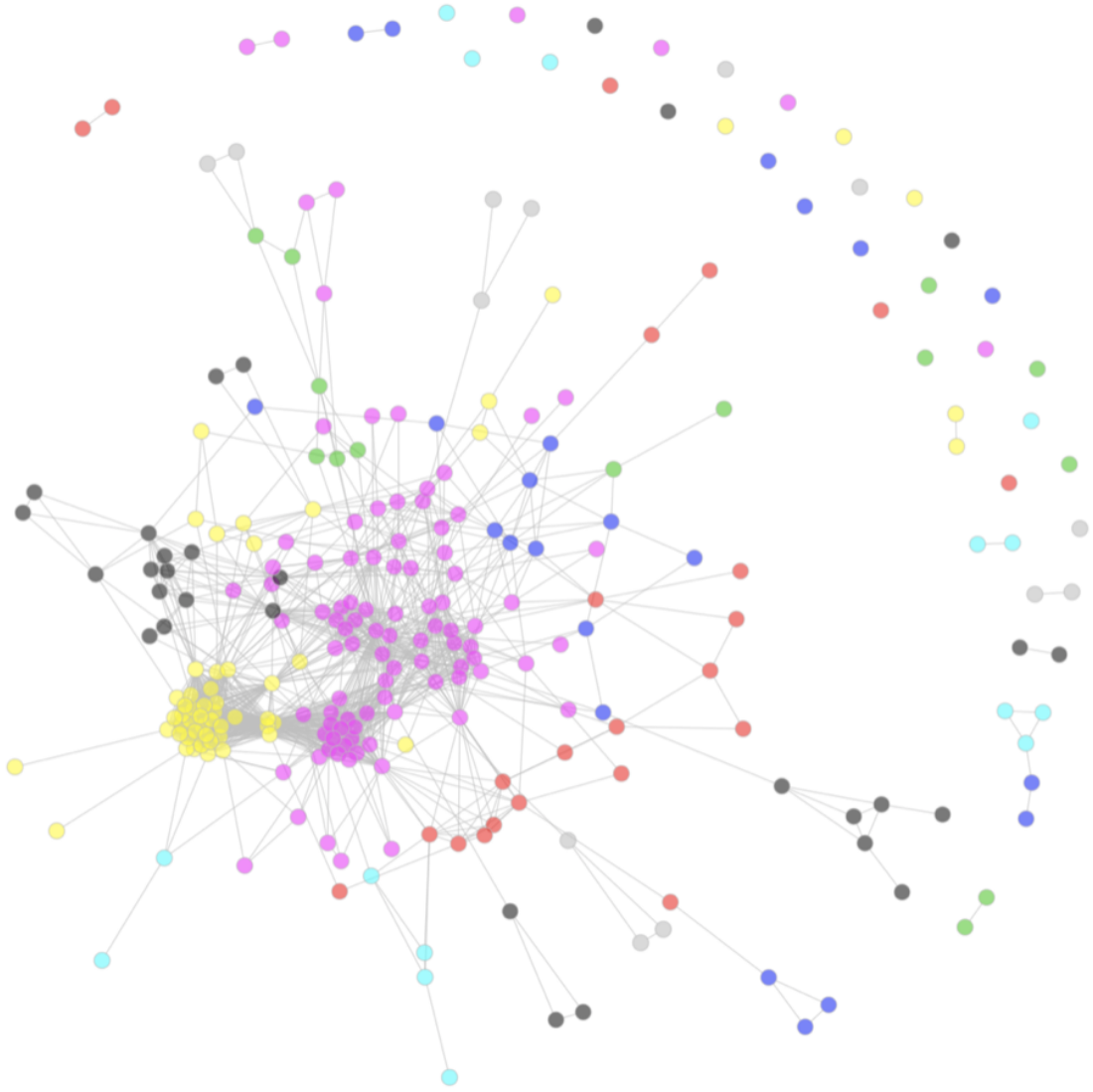


Figure 7: Founder networks based on Schools, Former Companies, and current Startup

Schools	Counts
Stanford University	27
University of Pennsylvania	17
Harvard University	17
Northwestern University	10
Massachusetts Institute of Technology	9
University of Michigan	8
University of Southern California	7
University of California+ Berkeley	7
University of Chicago	6
Georgia Institute of Technology	6
Yale University	5
University of Illinois at Urbana-Champaign	5

Figure 8: Most frequent Schools that Founders attended

0.4.2 Question 2: What are the most frequent schools or companies?

After answering question 2, it is natural to further see what exact schools and companies are hidden behind the networks and whether they have some common features. As shown in the table of school frequency (Figure 8), the schools with high frequency include Stanford, Pennsylvania, Harvard, Northwestern, MIT, Michigan, etc. They are all top schools in the US and in the world. So it is clear that education background is quite important to founders. On one hand, those who attended these selective shools already have talent in some areas. On the other hand, the school networks and resources are likely to help those founders to achieve success.

When further looking at the most frequent former companies, we can reach the same conclusion. (Figure 9) Google, IBM, Caktus, Accenture, McKinsey are all top companies in the world. Since we are focusing on the health care industry, it is not surprising that some founders would have health industry, consulting, and academic background. But the technology companies like google also have high frequency, which indicates that HealthTech is a great area with innovation and was favored by investors. I hope the frontier technology can make more people live a better life.

0.4.3 Question 3: What are the most important or popular skills to be a good founder? What are the founders good at?

This might be a question favored by those who also have entrepreneur ideas. I used Linkedin data to show some key aspects here. From figure 10 we can see that 64% of the founders have a graduate degree, mostly MBA or Doctorial Degree. An MBA background founder would have great business insights and rich network resources, while a doctorial background founder would enjoy higher technical barriers and core competence. However, there are also large amount of founders whose highest degree is bachelor. Those younger founders may have more creative thinking and sensitive to new technology and trend, while the elder founders would have richer industry experience. Due to the industry character, a science or business degree is more common among the founders.

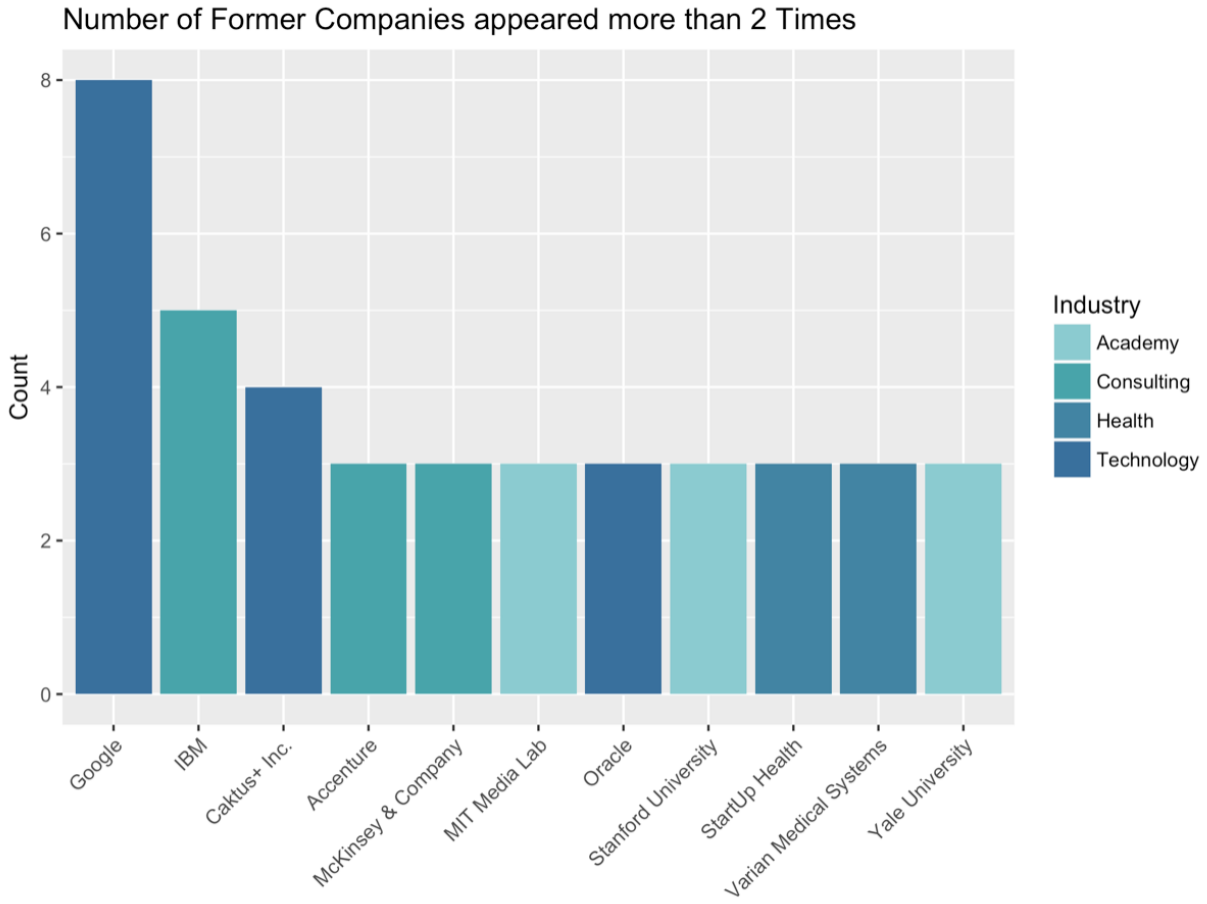


Figure 9: Most frequent Former Companies that Founders Worked For

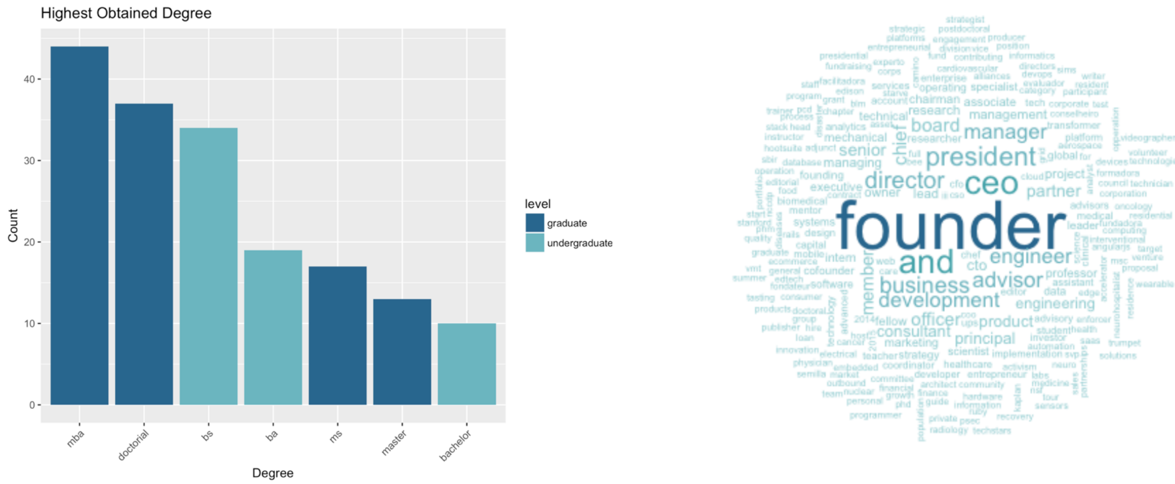


Figure 10: Founders' Highest Degree Obtained and Former Experience

Although the founders might have obtained diverse degree, it is clear that most of them have significant former leadership or business experiences according to the word cloud. 'And' here indicates that they might have multiple job functions at the same time, which indicates that being able to handle diverse tasks and solve different problems are also important features for a successful founder.

Looking at figure 11, this is an interesting combination because it shows that a good founder in the health industry is likely to have both business and technical background. The most frequently mentioned skills on Linkedin can be grouped as leadership, management, strategy, product development, marketing, public speaking, and industry understanding.

0.4.4 Question 4: How is their influence power on the social media? Do they have a lot of tweets, followers, retweets, likes?

The last question to answer is related to their social media influence. Nowadays social media has become one of the most important communication tools, and social media also provides rich data that worth mining. As can be seen in the former analysis, leadership skills (influence power) are crucial to be a good founder. I am here using Tweeter data to go beyond just labeling founders with certain traditional features, and try to see whether there's any pattern out there.

Here I compared the number of founders' followers, tweets, retweets, and likes (which could reflect the founders' influence power) with the funding round (which could reflect the startup's successfulness). (Figure 12, 13) Since a number of data points after round 3 are small, I only used the data within round 1 - 3. We can see a positive relationship here. The founders having received later rounds tend to tweet more and have more followers.

However, when we see the retweets and likes, the difference is not significant according to the median value. To reach a more robust conclusion, I do need more data points to minimize the selection bias and increase randomness.

But we have another observation here. There are quite a lot outliers of the count of retweets and

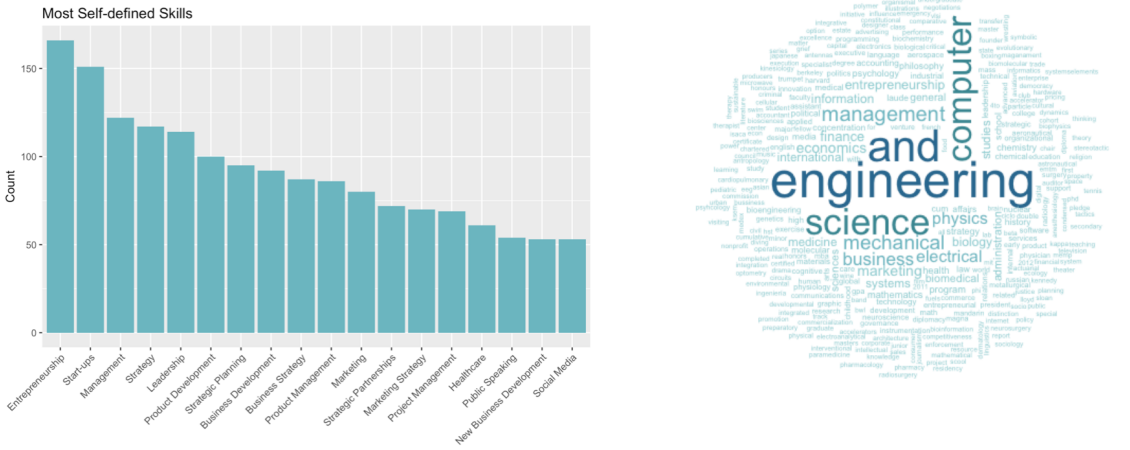


Figure 11: Most frequently mentioned Skills and Founders' Education Background

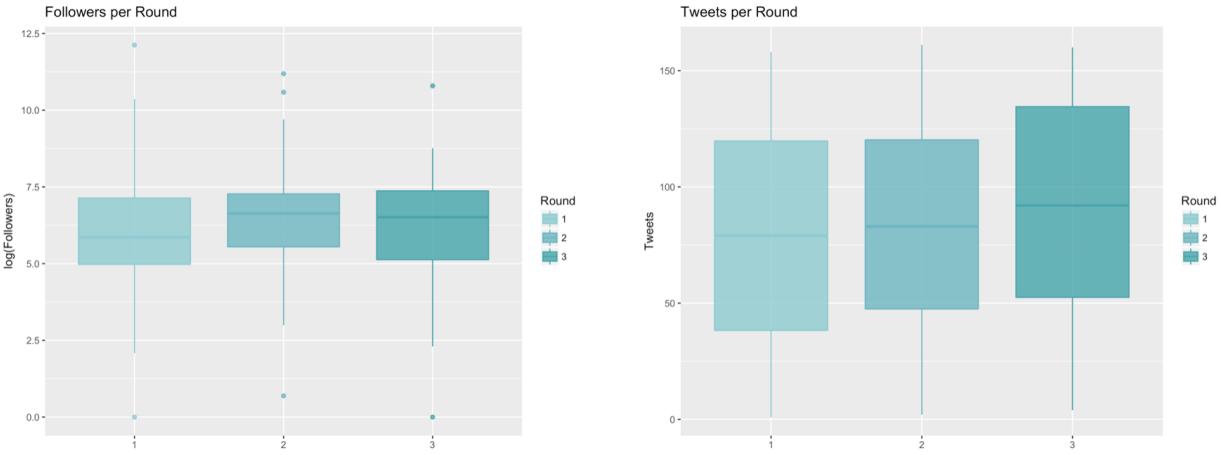


Figure 12: Most frequently mentioned Skills and Founders' Education Background

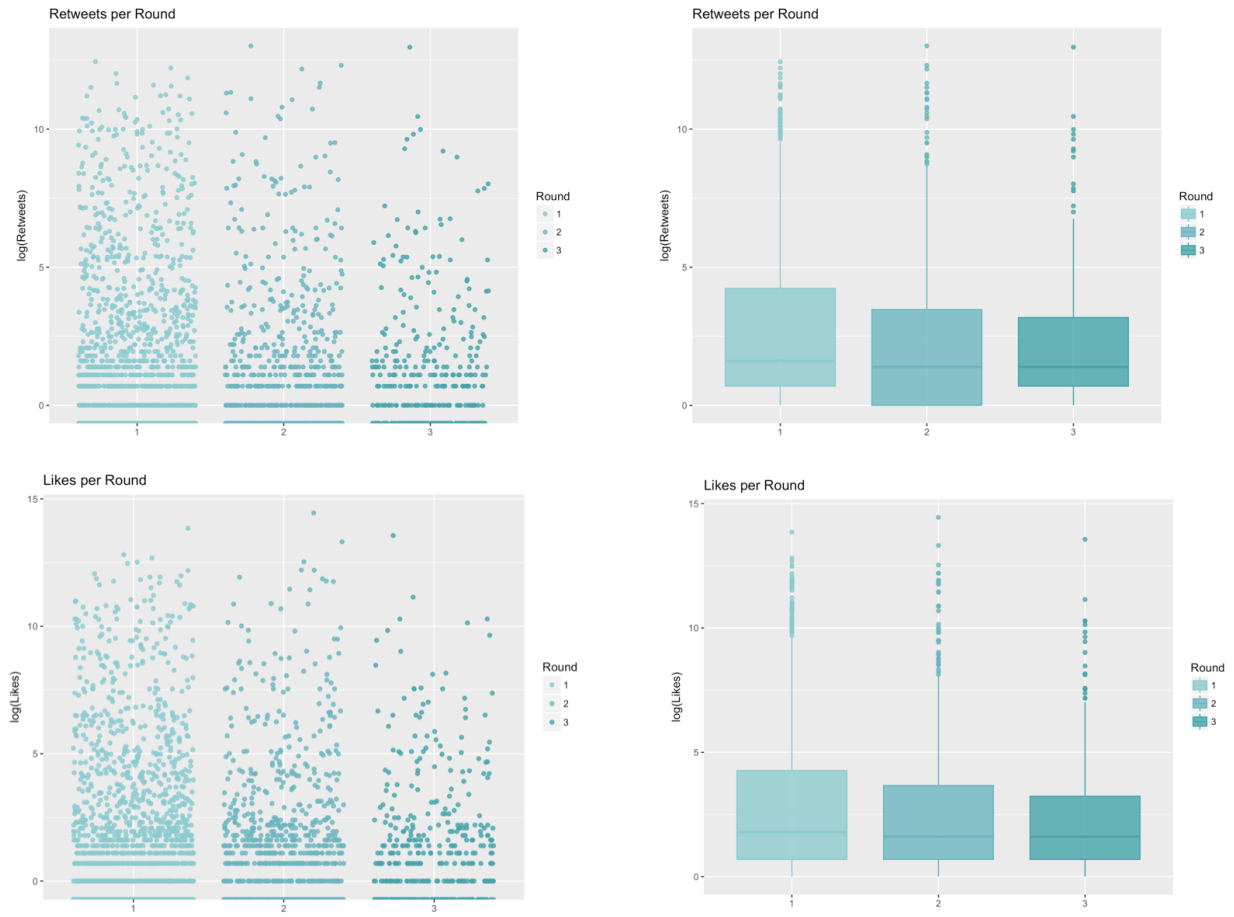


Figure 13: Most frequently mentioned Skills and Founders' Education Background

likes in the dataset, which means that the cascade of information does occur from those founders. This can also reflect their influence scope on the social media is large.