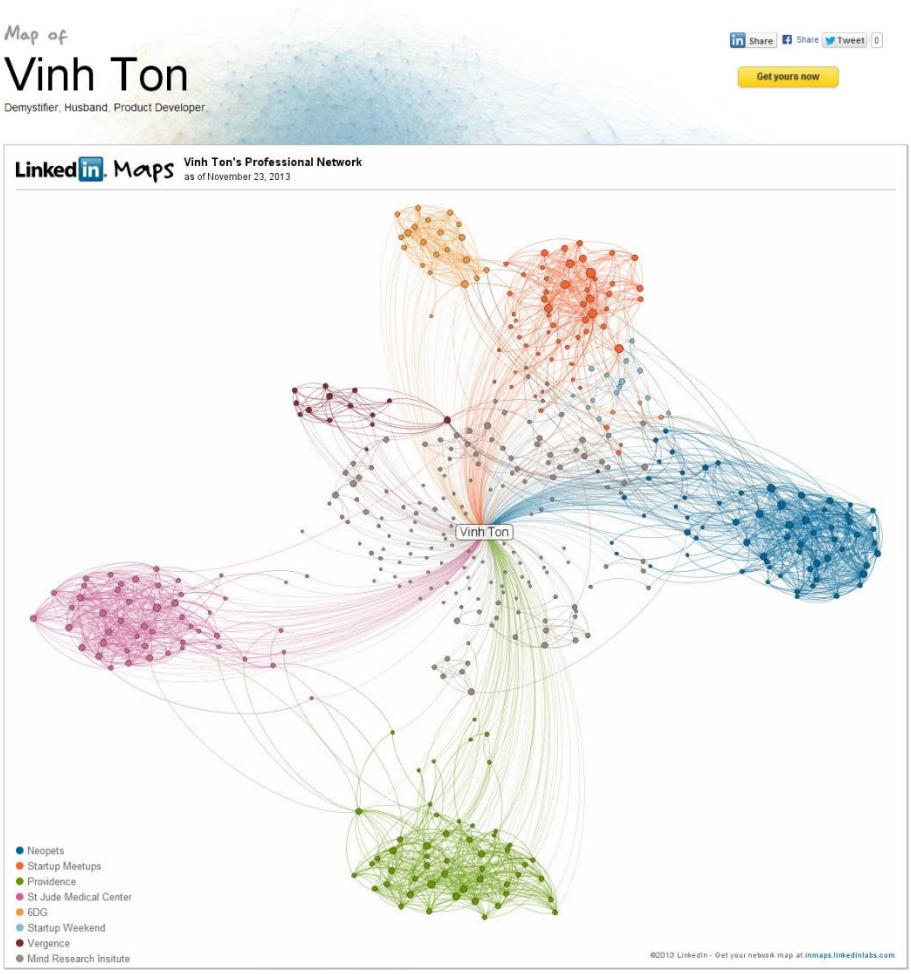
My Linked In Analysis

It is not how much but what communities and the position of a person in it



Contents

Introduction

Professional networks have a tremendous ability to influence the opportunities we have access to. The existing partnerships we have are defined as assets by the business savvy. Linked In has stood out as the premier online professional network. With Linked In's Application Programming Interface (API) and other technologies, there is a way to analyze our networks.

By analyzing my Linked In network, I hope to find a more strategic approach to leveraging it. I see it as analogous to financial analysts evaluating the different types of financial assets and how best to use it. This analysis will show that it is not enough to value a professional network by its quantity. What is more insightful are what the groups of people who have tight connections. This will be referred to as ***communities***. The analysis will show what communities exist in my network and how to find them in yours. Knowing the communities within your network will allow you to tailor specific content and approach if you want to spread information or influence. The analysis will also show that evaluating specific people in your network (***connection***) by how many connections (***degree***) is not adequate. More informative are where your connections reside in your network, also known as ***betweeness***. Also covered will be how intuitions about networks such as geography and profession are literally only half the story. Ultimately, going through this analysis will give a great introduction on how to understand a real life social network through quantitative means.

Concepts such as social networks have been undervalued and not even comprehended in the past. Only recent has technology and mainstream adoption made analysis of such networks available to a laymen like me. I think it is only a matter of time that the average person will demand analysis of their networks in the same manner that they seek financial planning advice.

This report has three parts:

- How to get the Data and format it for analysis (What Data and Method)
- Analysis of my Linked In network
- Conclusions and Discussion

Feel free to skip to your section of interest. I imagine this report will appeal to data miners, network analysts, marketers, and social researchers.

What Data

How

LinkedIn offers the ability to query their network if you sign on as a developer. It has a tutorial to get started. LinkedIn also has the inmaps app. You can only visualize the network with inmaps, not do analysis on it such as is available in Gephi. There are others who have analyzed their Linked In but I was not able to get their code to work in regards to defining the edges. This led me to write my own code in Python Notebook. The previous work posted on dataiku has broken links so no results are viewable. It also seems limited to what connections are while Linked In's API provides other attributes such as heading and geography.

Criteria

My 1st degree connections made up the nodes and the connections between each other made up the edges. Due to privacy limitations, this was the only option available. It is probably the only feasible approach as far as computing goes. LinkedIn limited my query to 500 or less a day. Some of my connections had privacy options on so their connections are not included.

Method

If you have access to your own Linked In data including edges, then you can skip straight to the Gephi portion. If you are familiar with Linked In and Python, then you can go straight to the code I wrote to extract your Linked In network data and convert it to a graphml file that works for Gephi.

A word of caution, this might get frustrating if you are not comfortable with programming. Even as an experienced programmer, I encountered several annoying issues such as the Linked In throttling and international characters. This will be an overview of the whole proess and not detailed technical implementation of the API, Python Programming, or Gephi analysis.

Get LinkedIn Developer Authorization

First, you must get your authorization keys and tokens from Linked In by signing on as a developer. This is relatively easy as all is needed is to sign on to the Linked In developer site.

Install Python Notebook

You can use another language if you prefer as Linked In allows it. I went with Python because there was existing work from Matthew Russell (author of Mining the Social Web), Dataiku, and Linked In's tutorial. Each of these sources has pieces of the puzzle but not one was fully operational nor was dedicated to analyzing Linked In connections in detail. Russell came close but only introduces clustering.

I was most impressed with Matthew Russell's virtual machine setup available on github. This is perfect for those who want to hack their way to analyzing the Linked In data. I have no interest in learning the intricacies of Linux or the other pieces that make up the process. It is ridiculously simple to use Russell's process for setting up an environment where you can code in Python Notebook within the hour.

Run Python Code

Assuming you have your Linked In keys and a Python environment up and running, then you can get my code on Github. You have to fill out the myLinkedCodes.xml with your own keys. Save the xml in the same directory as your python code. Later updates may change this to put data in its own directory.

I cannot stress enough the importance of understanding how Linked In limits the number of api calls a developer can make to its network within 24 hours. This is also known as throttling. Because of this limitation, you will most likely run more than one session spread out between days. You might get away with it if you have a network less than 400. This does make things more complicated but the code takes it in to consideration. All you need to do is manually adjust two variables:

- **THROTTLE_LIMIT**: this defaults to 200 because our limits will vary. You can probably push it up to 400 if you haven't made any api calls to Linked In all day. I don't recommend it though as you are better off saving what you have and running multiple sessions.
- **LAST_NODE_ID**: this is the id of the last processed node before stopping. You get this from the output of the "Read in GraphML and finish adding edges" code cell. It tells the code where to pick up from rather than starting from the beginning of all your connections.

Run Gephi

Having gone through all the steps, you should now have a graphml file that is suitable for Gephi. There is a quick start tutorial from Gephi that I highly recommend.

For my network, set up the labels as the first 15 characters of the heading. This allows for anonymity but still provides informative identifiers. A combination of Force Atlas 1 and 2 was used to separate the nodes by their connectedness.

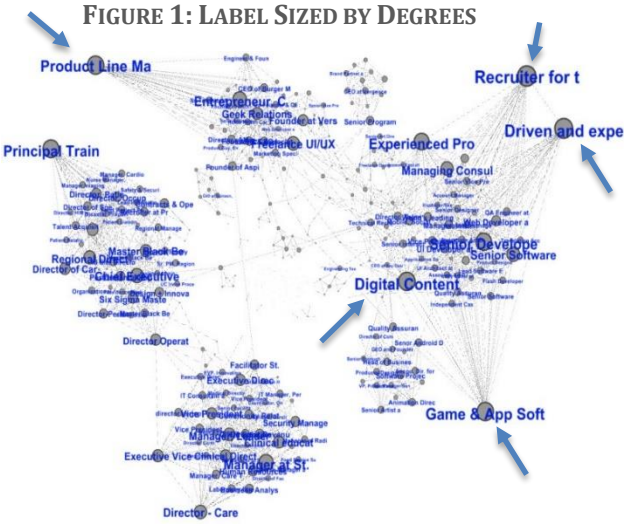
Data Analysis

Metrics

Average Degree

There is an average of 7.8 connections for those in the network. Figure 1 shows the nodes and labels sized by the degrees. With arrows indicating the top 5 people with over 34 connections within the network. It is no surprise that the Recruiter is the highest with 39.

FIGURE 1: LABEL SIZED BY DEGREES



Average path length

It takes an average of 4.8 connections to reach people throughout the network.

Betweeness

Betweeness is one of the most practical concepts I learned from the University of Michigan's MOOC on Social Networks Analysis. Consider Figure 2: Triad and Figure 3. Both smiley faces have a degree of 2, or 2 connections. However the green smiley face in Figure 3 is in a "broker" position where it is critical to bridging connections.

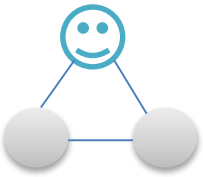


FIGURE 2: TRIAD

FIGURE 3: VALUE OF BETWEENESS

Now consider my real life network and the drawback of merely counting how many connections a person has. Figure 4 shows my network rearranged where the size of the nodes and labels indicate the person's betweeness rather than their degrees. Compared to Figure 1, it is clear that the people who just

knew a lot of people are not the same people who play vital positions in bridging the different groups. I will keep these people in mind as key components of facilitating information across communities.

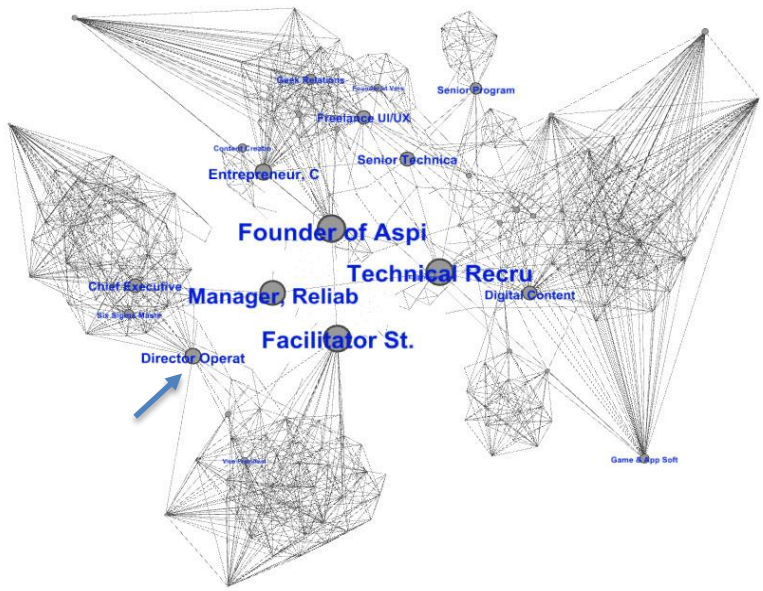


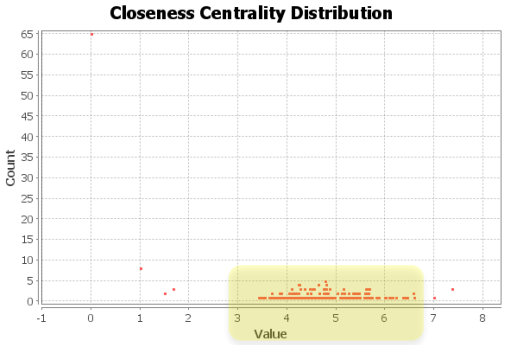
FIGURE 4: SIZE BY BETWEENESS

Another point of interest is who stood out as brokers. The recruiter and gregarious personalities were obvious. However the Director indicated by the arrow does not fit the stereotype.

Closeness

This is the concept how close a person is to the action. Like how well used a road is within an urban network. With the closeness distribution in Figure 5 it is possible to hypothesize that information can spread through the network in around 5 hops.

FIGURE 5



Communities

Before SNA

Before I knew about Social Networks Analysis, I would make intuitive guesses of what formulates a community. For example I would assume shared attributes like profession would dictate communities.

As we see in Figure 6 and Figure 7, half of the network is dominated by a few industries and they do form some of the communities. The rest is quite diverse.

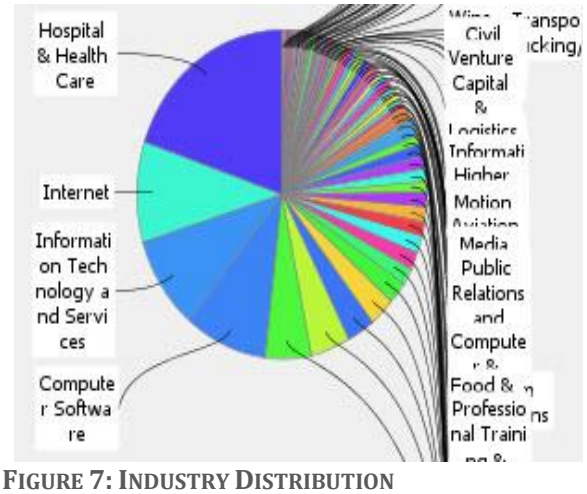


FIGURE 7: INDUSTRY DISTRIBUTION

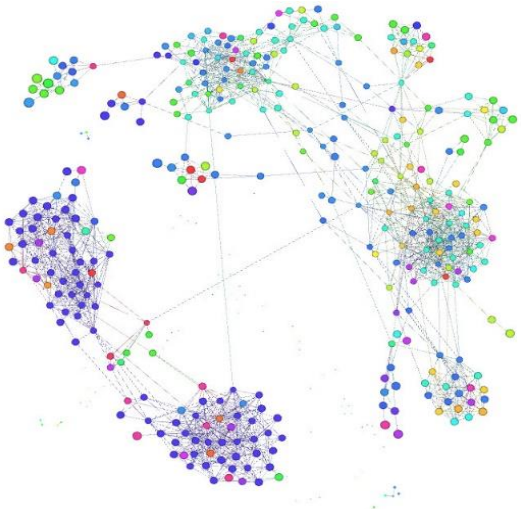
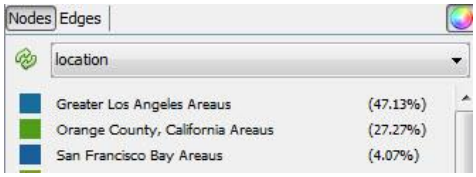


FIGURE 6: COLOR BY INDUSTRY

Next I would make guesses about geography as people who are in the same vicinity should form communities. Figure 8 shows most of my network is either in LA or Orange County.



However, this is not that informative. I don't need pretty pictures to tell me that there are communities formed around where people live.

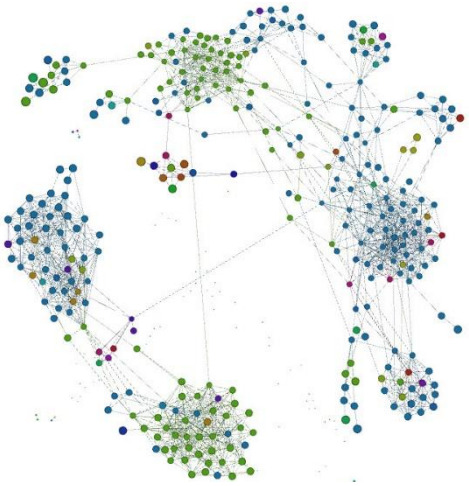


FIGURE 8: COLOR BY LOCATION

After SNA: Quantified Communities

I am extremely impressed about how the numerous approaches to identifying communities has matched what I know about my personal network. There are a mixture 2 or 3 approaches that I will cover: Linked In's Inmaps and Gephi's Force Atlas combined with Modularity. I think this is a mixture because they all probably use similar algorithms or some mix of it.

Linked In's InMaps

Linked In's Inmaps is able to separate the communities by what I knew as my professional communities as shown in Figure 9.

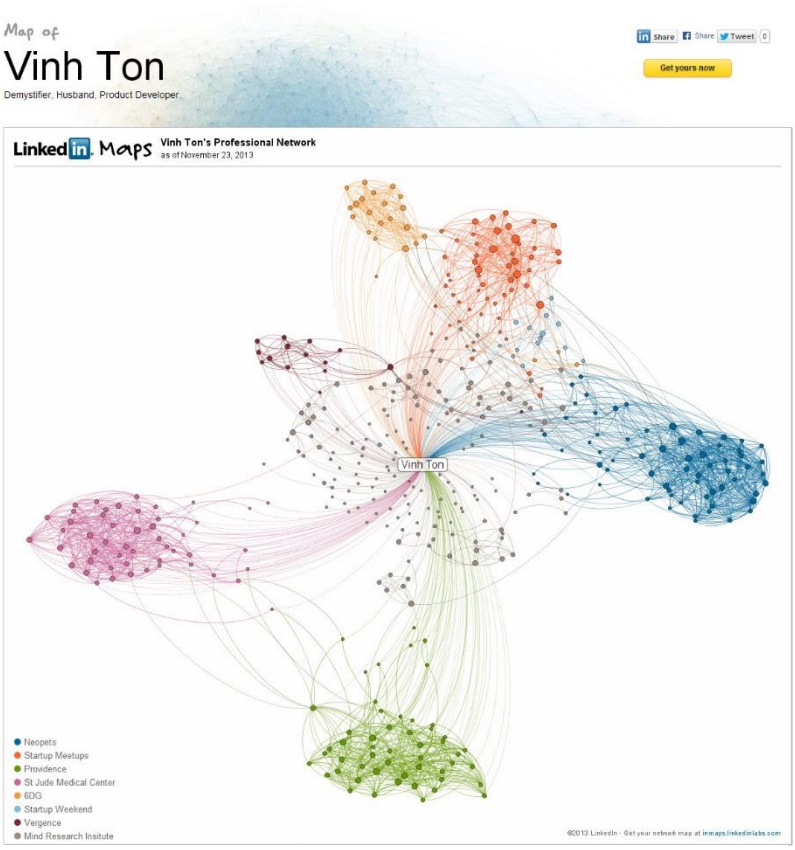


FIGURE 9: LINKED IN'S INMAPS

Gephi

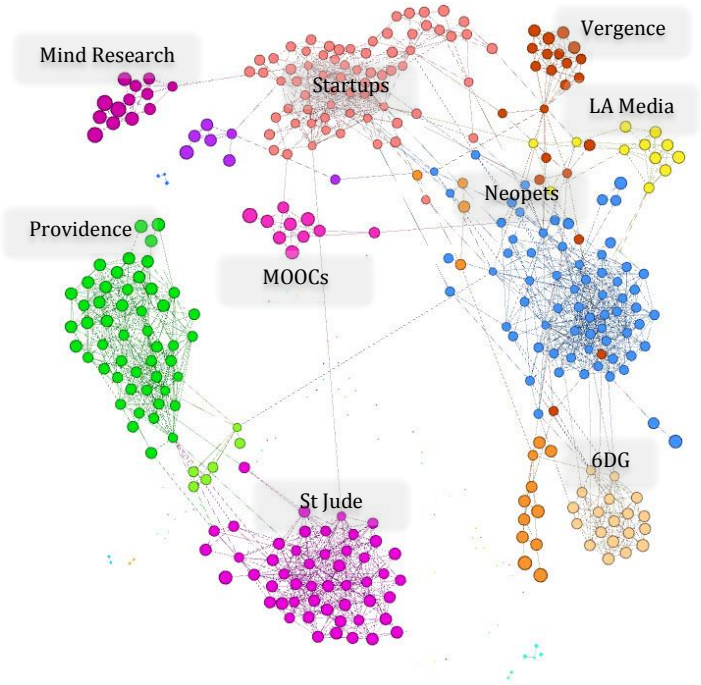
The Force Atlas algorithm combined with the modularity classification gives a more insightful picture of the communities within my network as shown in Figure 10.

Gephi's approach is able to detect more distinct communities than In Maps. I have validated those are real communities by labeling them as I know them.

Conclusions

This analysis shows quantitatively who the key players are in my network by using the measurement of Betweeness. This gave insights to properly evaluating how critical a person is in my network where intuition alone was not adequate. It was also useful to distinguish the communities within my network. Finally the process informed me of the drawbacks and advantages of available tools. Since I used my own network, I can validate the value of the quantitative analysis as true.

FIGURE 10: FORCE ATLAS AND MODULARITY



Discussion

Limitations

It was quite a pain to develop my own program over a period of a month to extract the data and convert it to a useful format. Given that others have done something similar but the programs are no longer operational or there has been significant revisions shows that data mining will not be a consistent process. I expect my implementation will also grow outdated.

Another limitation is that I only have access to my own network. The community detection validated what I already knew about my own network. This will be more useful for an unfamiliar network.

Future Studies

It would be much more powerful if there is access to data that defined the level of influence a node has. The data does exist. Examples include when the person posts information, how often is it clicked or shared. Unfortunately, I either can't or have not figured how to access that kind of data. If such a measurements of influence was available, it can be used as weights for the nodes or edges. This is will give a very powerful visual of the flow of influence.

To use terminology in the book The Tipping Point, betweeness is a perfect analogy to Connectors. If we can measure influence, we can identify the Mavens and Salesmen. This would allow for numerous applications of influence and information dissemination. From commercial applications of marketing to life saving persuasions of getting immunization, there is so much potential.

About the Author

I'm a passionate Social Entrepreneur and am always on the look out to connect with good people who can execute and share my passion to make the world better. Connect with me at www.mindgnosis.com if interested. I think SNA will be a critical component to the evolution of our society.