Relationship between Tier in Supply Chain and Industry Classification of Company

Hao Fu

Shanghai Jiao Tong University, SJTUFuHao@sjtu.edu.cn

Abstract

The aim of this report was to study the relationship between tier of a company in supply chain network and its industry classification. We constructed the supply chain network based on the supply relationship of quoted companies and calculated the average tier of every company in the supply chain. After that we used the NAICS classification of every company to find the different tier each industry occupies and data visualization technique to make this relationship more clear. The report concludes that companies' industry classification have strong relationship with their position in the whole supply chain network, and the traditional industry classification to some extent is outmoded. It is recommended that we can form a new industry classification based on company's tier.

KEYWORDS

Supply chain network, Industry classification, Data visualization, BSF, DFS

Introduction

There has been a massive research on the supply chain to manage company risk and predict firm performance, but the focuses are mainly on local supply information of specific firm, not on the global supply chain network property. Also as time goes on, many manufacturing firms like IBM have developed to service companies, but the industry classification system, like NAICS(North American Industry Classification System), does not consider this change.

So in order to resolve the aforementioned problem, we try to construct the whole supply chain network of the quoted companies and calculate the average tiers of each

company. We also use the NAICS of each company to find the potential relationship with companies' tier. We also plot the supply chain network to make this relationship more clear.

Methods

Data preparing

In this research we used the first dataset that has the suppliers and customers information in 2003.04 to construct the supply chain graph. Since each quoted company has an distinctive CUSIP ID, we use this ID to identify different companies. The graph we constructed called 'graph1' is a directed graph, the direction is from the suppliers to the customers. We also construct a reversed graph of graph1 called graph2. The direction of graph2 is from the customers to suppliers. The standard CUSIP ID has 9 digits, but the last 3 digits are useless in this research. And there are 8-digit CUSIP ID and 9-digit CUSIP ID in this dataset, so we just cut the first 6 digits to represent the company.

We also used the second dataset that has the NAICS information of each company to construct a dictionary that can quickly return the NAICS of any given company. There were different lengths of CUSIP ID in the dataset, from 6 to 9, so we neglected the records that have CUSIP ID shorter than 8 digits because we can't use such CUSIP ID to identify only one company. If the CUSIP ID only had 8 digits, we added an extra '0' in the beginning of the CUSIP and cut the first 6 digits. As for the 9-digit CUSIP ID, we just simply cut the first 6 digits to indentify the company.

The details of the process of data preparing can be found in the source code "scn.py" and "naics.py".

Another important process is to remove loops in the supply chain graph. In the supply chain network companies such as manufacturing provide products to companies of service industry, and also the companies of service industry provide finance support and service to manufacturing company, so the loop in the supply chain must be very common. But since we interested in companies' tier in supply chain network, the loop can be a confusion because the company in one supply chain can be count for more than one time, so we need to first remove loops in the graph. We used DFS algorithm to remove the edge that may cause loop, and to make this faster we optimized the code.

Data processing

Since we have the prepared data from the abovementioned process, we can start to calculate the tier of each company. The supply chain network is such complicated that each node, that is each company, in different supply chain has different tier. So, in order to fully describe the position of a company in the network, for each company we calculate its tier in different supply chain. We define the tier of nodes with zero in-degree as tier 1.

$$Tier(d_{in} = 0) = \quad (1)$$

As for node with non-zero in-degree, the tier can be defined recursively: the tier of a non-zero in-degree node is higher than the tier of previous node in this supply chain path for 1.

$$Tier_j(d_{in} \neq 0) = Tier_i + \quad (2)$$

So we can calculate the tier of a node in different supply chain path use abovementioned definition. This is one of the position of a company in the supply chain network. To estimate the overall position of a company, we need to determine the average tier of each company.

$$Tier_{v,e} = \frac{\sum_{i=1}^{n} Tier_i}{n} \quad (3)$$

Now we have two tables: the first one is the industry classification of each company, and the second one is the average supply chain tier of each company. Using this two tables, we can see if there any correlation between this two variables.

The details of data processing can be found in the source code "process.py".

Data visualization

To make the relationship between the industry classification and supply chain tier more clearly, we used several data visualization tools to show the result.

The first one is 'Gephi'. 'Gephi' is the leading visualization and exploration

software for all kinds of graphs and networks. Here is an sample graph made by Gephi.

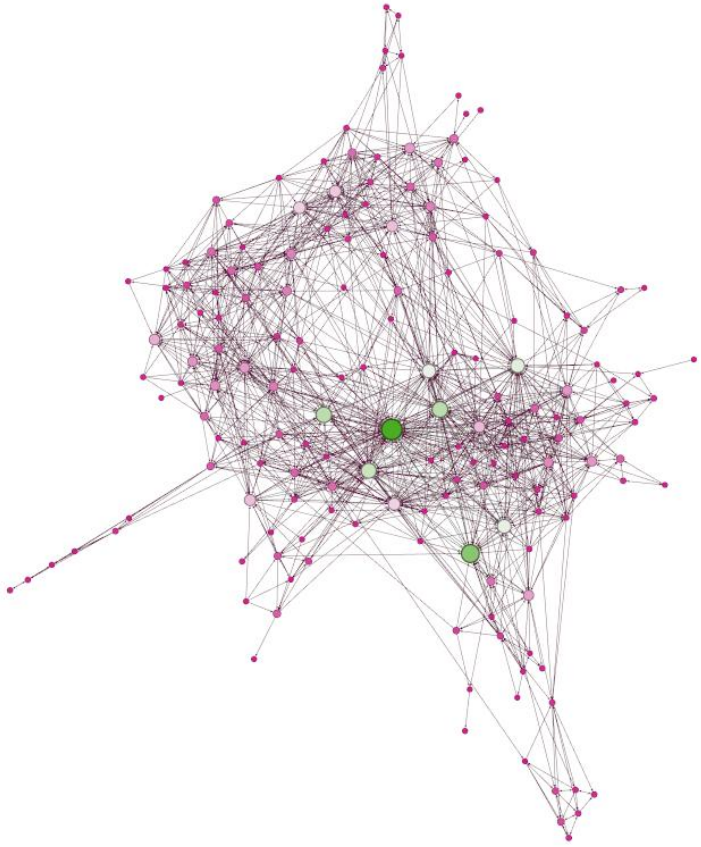

Figure 1 Gephi Sample Graph

There are several layouts for different kinds of graph. The layout we used in this research called DAG Layout, which is a simple layout for directed acyclic graphs (DAGs). The nodes are arranged in discrete layers so that the edges will always point downwards (if no loop exists). The nodes are arranged as far to the top as possible, minimizing the number of layers used. The horizontal layout is done by assigning the nodes to discrete slots in each layer (the biggest layer defines the number of available slots for each layer). While running, slots are chosen randomly and swapped if this would make the edges shorter to generate a more compact graph. These optimizations are only local and will not generate an overall optimal layout. Here is a sample graph.

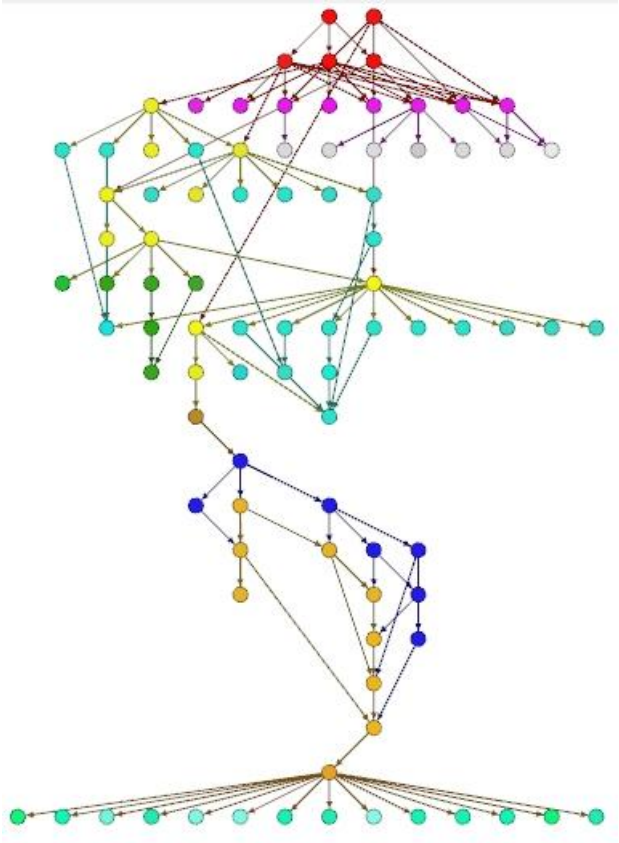

Figure 2 Sample Graph of DAG Layout

We used DAGs layout draw the supply chain graph, in which all the loops have been removed. For company of different NAICS we used different color to paint. The larger the first two digits of NAICS, the darker the node in the graph. Also for node with larger in-degree and out-degree, we set it to be larger.

Another visualization tool we used is Tableau, which produces a family of interactive data products focused on business intelligence. We used Tableau to draw the stacking graph, which can show clearly the relationship between section and entirety. Here's a sample.

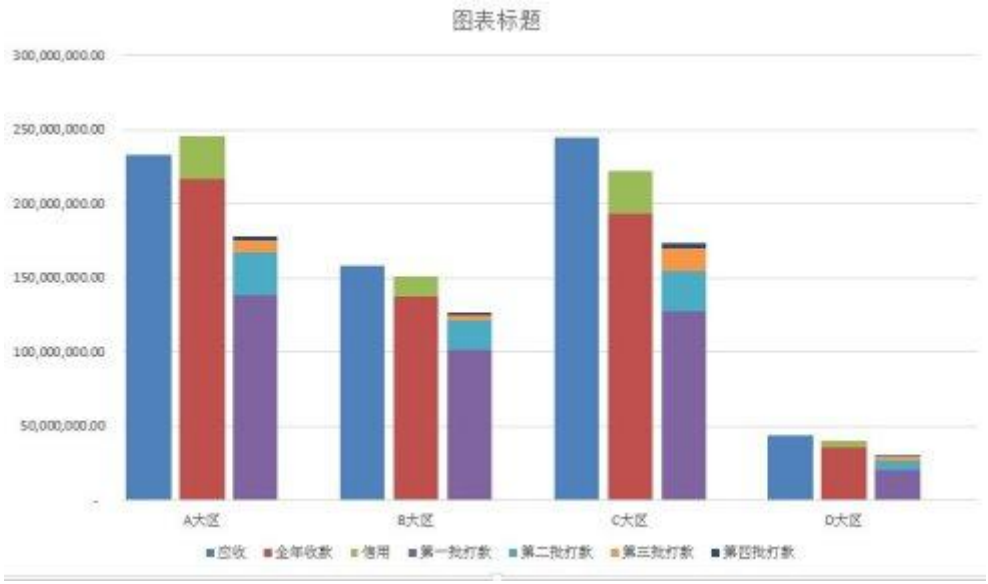

Figure 3 Sample Stacking Graph

We used this kind of graph to illustrate the NAICS's distribution in different tiers.

Results

After the abovementioned processes, the DAGs graph is constructed as you can see below.

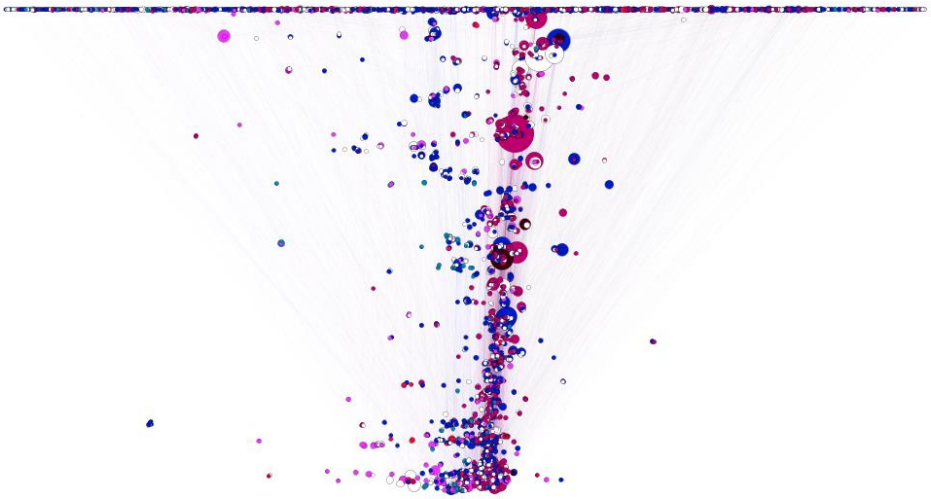

Figure 4 DAG Layout Graph of Supply Chain Network

Also, the stacking graph of NAICS's distribution in different tiers is here.

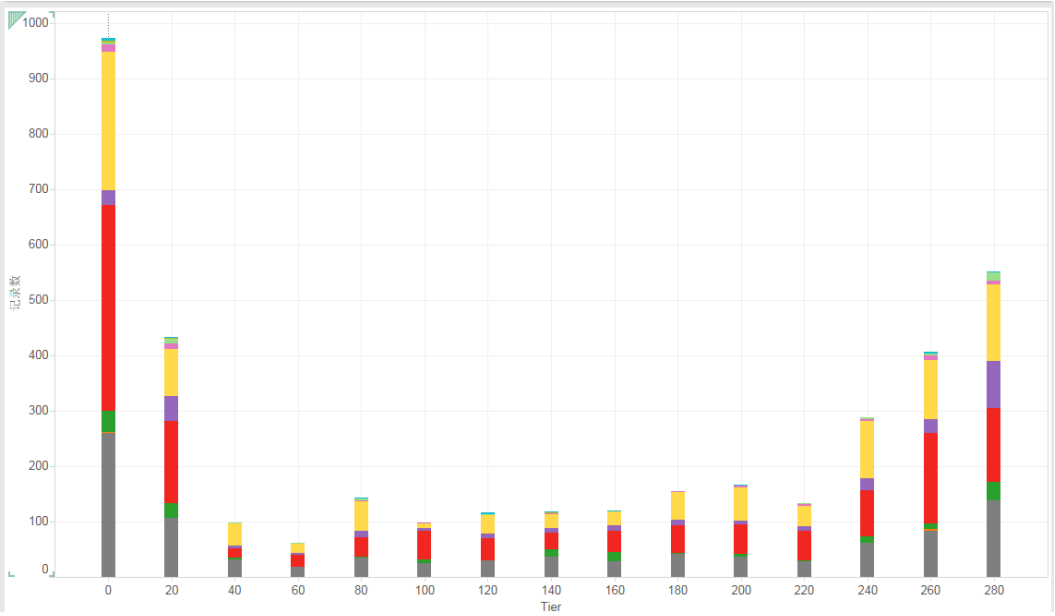

Figure 5 Stacking Graph of Different Tiers' Industry Distribution

Discussion

For the DAGs graph constructed by Gephi, we can find that although companies of different industry distribute from the low tier to the high tier, the high tier concentrate more dark nodes. Since dark node means high NAICS, and the NAICS ranks from primary industry like agriculture to the secondary industry like manufacturing and tertiary industry like servicing, we can conclude that most of the companies of tertiary industry are located at the higher level of the supply chain and the companies of secondary industry are more likely to be located at the lower level than companies of tertiary industry in supply chain. Another interesting finding is that most of the companies are at the lowest tiers and the most highest tiers, while the companies in the middle tiers are so few, and also most of the companies with high degrees are at the middle tiers.

To make this more clearly, let's see the second picture, the stacking graph. We can see over 2500 companies' average tiers are lower than 40 or higher than 240, while the whole company number are 3863. This also means that the companies at the middle of the tier receives the raw materials from companies in low tier, and after processing them supply the products to the companies in high tier. And the company at middle tier often has much more supplier and customers, so this sort of companies are very important for the whole supply chain network because they intercommunicate the downstream industry and Upstream industry.

There are also some imperfect in the results, such as although we can see companies of tertiary industry are concentrated in the higher tier than of primary industry and secondary industry, there are still much in the lower tier. What we can do more is to improve the method of calculating the tier. Another method is to normalize the tier, that is to set the node with zero in-degree in 0 tier, and the node with zero out-degree in 1 tier, each company's tier is larger than 0 but smaller than 1. Because the longest path in supply chain is over 500, another company at the highest tier, that is with zero out-degree, might be at above 20 tier. This may cause some problem when the distribution of tiers is too wide.

Conclusion

From the aforementioned discussion, we can conclude that the industry classification do have strong correlation with the company's position in the whole supply chain network, the companies of high level industry are often located at higher

tier of supply chain than of low level industry.