3.1.3 Explanatory Analysis

Upon the completion of the construction of the competitors network we proceed by investigating the main characteristics of the network. The competitors network is a directed graph consisted of $|V| = 2.327$ nodes and $|E| = 3.429$ edges. Recall from Section 1.2 that for a directed graph the number of all possible links between its nodes is calculated as $N \cdot (N-1) = 5.412.602$. To be more precise the graph density $D = 0.00063$. We therefore conclude the competitors network is very sparse as it can be seen in Figure 3.2 below.

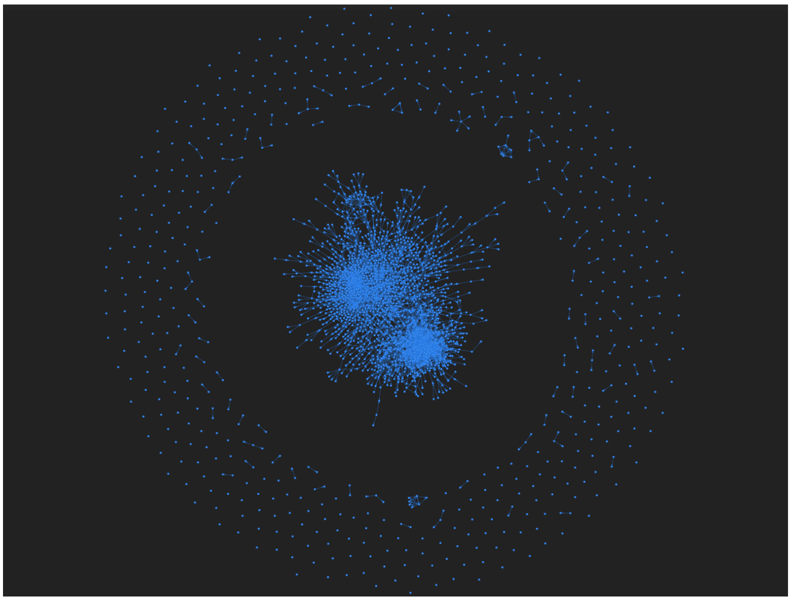


Figure 3.2: The Competitors Network created by parsing SEC 10-K filings and capturing companies mentioned as competitors with the use of named entity recognition.

Indeed one can easily see that there are many individual components of different sizes with the majority being of size two. We calculated the *weakly connected components* of the graph which do not consider whether the direction of the edges. We provide the following bar plot to have a more clear picture since the large size of the graph makes it difficult to visualize all the (weakly) connected components that consist the network. It can be seen from the Figure 3.3 that the network consists of a one large component with 1.708 nodes while the size of the rest components is significantly smaller (ranging from 1-7 nodes). In order to get a sense of the size of the network we calculate its *diameter* which is simply defined as the largest path between any pair of nodes. However since we argued that the network consists of disconnected components we choose the largest one because it is safe to assume that there is no larger diameter for the other smaller components. The network diameter of this network's largest component is 16; the greatest distance between any two nodes far away from each other.

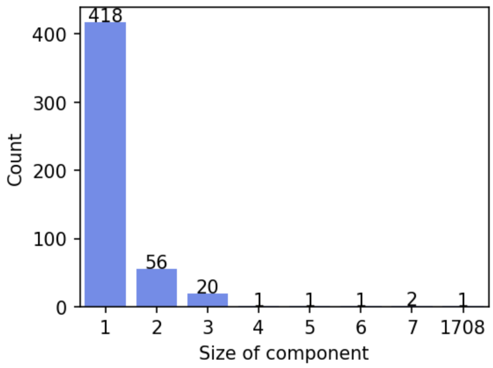
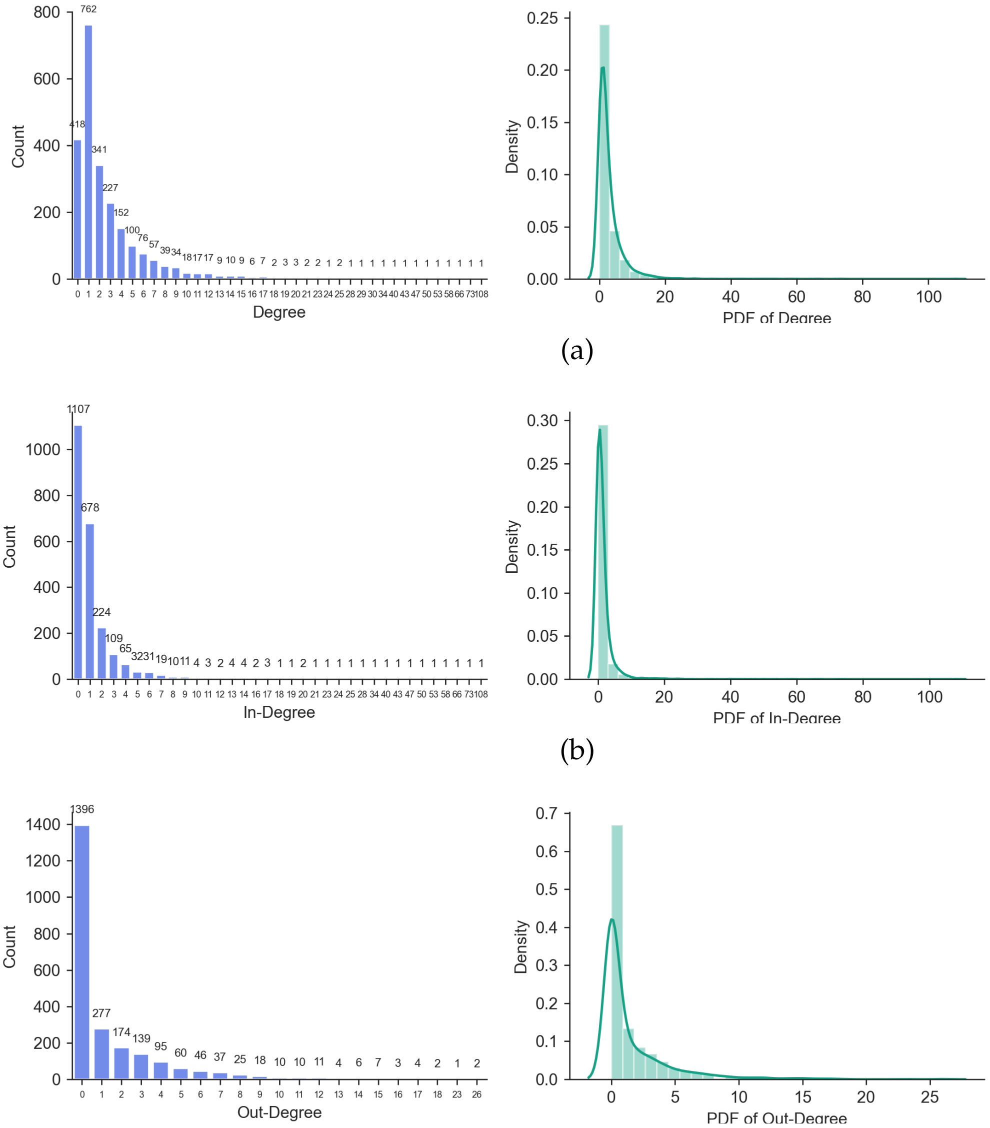


Figure 3.3: Frequency of individual components of different sizes

The next structural calculation involves the concept of *triadic disclosure*. As seen in Section (1.2) the two most common metrics for triadic disclosure are the *clustering coefficient* and *transitivity*. By applying equations (1.9), (1.12) we obtain $\overline{C} = 0.025$ and $T = 0.28$ respectively. Note that it should not be surprising that the transitivity of the graph is much higher than its density. Since the graph is very sparse there are less possible triples to form in the graph, or mathematically, we see that the number of possible triples appears in the denominator of equation (1.12) for transitivity.

Next we proceed by searching for the most important companies in the competitors network. The degree is the simplest and among the most common feature for measuring the importance of a node and it is simply calculated as the sum of its edges. Since the competitors network is a directed graph we also consider the in and out-degree which is similarly the sum of incoming and outcoming edges respectively. The average node degree is 2.95 while the average in and out degree are equal to 1.47. In addition Figure 3.4 below displays the degree, in-degree and out-degree distributions. Clearly the degree distributions are right skewed which means that the majority of the in or/and out degrees are quite low and we rarely observe nodes with high degrees.

In accordance to our discussion in Section (1.2), real directed networks often consist of many different connected components of various sizes. The same holds for the competitors network and because of this phenomenon the majority of nodes have nearly zero eigenvector centrality scores. Interestingly, a similar picture corresponds to all the rest of centrality measures we tested to evaluate the node importance, in order to find the most central nodes while looking at them by different scopes.



(a)

(b)

(c)

Figure 3.4: Barplots (left) and probability distributions (right) for in-out degree (a), in-degree (b) and out-degree (c) for nodes in the competitors network.

The top 10 nodes with respect to six different centrality measures are summarized in table (3.1) below. Notice that the vast majority of nodes with high in-degree also had high Katz and PageRank scores. Recall that these measures are based on random walks over the network, each with some extra modifications (see Section 1.2). If a node has many incoming links then the random walk is very likely to visit that node. On the other hand, nodes with many out-links in the context of competitors is an indication that the

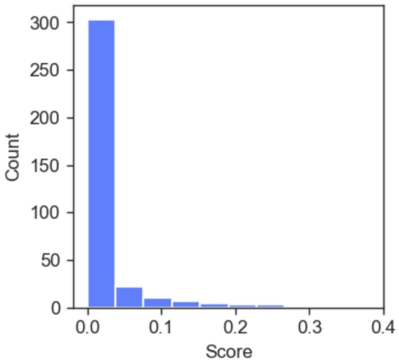


Figure 3.5: Histogram of eigenvector centrality in the competitors network. Frequency is given on the y-axis while x-axis displays the centrality scores.

underlying company might not be very competitive. This explains why the companies with high-out degree do not appear to have Katz and PageRank centrality.

	In Degree	Out Degree	Eigenvector Centrality
1	**Pfizer**	Durect	**Gilead Sciences**
2	**Merck**	NantKwest	Allogene Therapeutics
3	**Facebook**	Gristone Oncology	**Pfizer**
4	**Microsoft**	**Mustang Bio**	Atara Biotherapeutics
5	**Amgen**	Curis	Precision Biosciences
6	**Apple**	Lipocine	**Mustang Bio**
7	**Johnson & Johnson**	Agenus	**Merck**
8	**Gilead Sciences**	Ziopharm Oncology	Sangamo Therapeutics
9	Abbvie	Gossamer Bio	Crispr Therapeutics
10	Biogen	Parsons	Abbvie

	Betweeness Centrality	Katz Centrality	PageRank Score
1	**Mustang Bio**	**Pfizer**	**Facebook**
2	Bellicum Pharmaceuticals	**Merck**	**Pfizer**
3	Gristone Oncology	**Facebook**	**Merck**
4	Agenus	**Amgen**	**Microsoft**
5	NantKwest	**Microsoft**	**Johnson & Johnson**
6	Ziopharm Oncology	**Johnson & Johnson**	**Apple**
7	Hercules Capital	**Apple**	Moody's
8	Fortress Biotech	**Gilead Sciences**	**Amgen**
9	Solar Capital	Abbvie	Medtronic
10	BridgeBio Pharma	BIOGENiogen	Walmart

Table 3.1: Top 10 rankings of different importance measures. The companies with the highest centrality scores with respect to at least three different metrics are highlighted in **bold** font style.