CS224W Project Report: Characterizing and Predicting the Economic Network of Corporations

Chung Fat Wong (cfw20@stanford.edu)

code: https://github.com/cfw20/cs224w

1 Introduction

Corporations are not independent entities, but are linked to each other through many types of relationships. Examples of these links might be the relationship between a supplier and a customer, in which case the link is explicit and contractual. On the other hand, two corporations might be classed in the same industry in which case the link is more subjective. In this paper, we focus on a network model of corporations based on the former type of relationship. Specifically, the nodes of the network represent individual firms, and there is a directed edge from one firm to another if the former (supplier) sells goods to the latter (customer).

Supplier and customer links constitute one aspect of the supply chain network of corporations, and hence are important to study given that they form the backbone of many industries. Similar to recent studies on supply network topology, for example [1], we explore this graph using a variety of network analysis techniques. In addition, we use historical stock price movements of connected nodes as a novel empirical measure of the usefulness of the theoretical models. For example, we identify structural roles in the supplier and customer network to determine whether some customers should have an out-sized influence on the stock price of their supplier firms. We also detect community structures within the network to identify groups of corporations which should have correlated stock price changes.

The simplest depiction of a supply chain is a linear chain consisting of six nodes, representing each of the following entities: Raw materials, Supplier, Manufacturer, Distributor, Retailer and Customer (Fig. 1). In the modern world, this linear chain model is insufficient because two main factors: (i) the complexity involved in manufacturing today's products means that each manufacturer could have several suppliers upstream as well as itself supplying to several nodes downstream and (ii) modern supply chains do not purely consist of goods flowing from one node to the next, e.g. two corporations might be linked by an agreement to co-create patents for product design. By focusing on supplier and customer links only, our supply network model captures most of the complexity of (i), whilst avoiding the added complications of (ii) since many of these are not easily quantifiable.

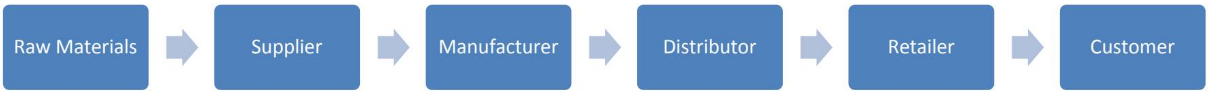


Figure 1

2 Related work

Although there is a large body of work on supply chain management, the study of the topology of the supply chain networks themselves first began around 2000. Since then, researchers have studied the structure of supply chain networks at various levels, going from the whole-network level down to node level.

At the whole-network level, the degree distribution of supply chain networks has been widely studied. For example, in [1], it was shown that the degree distribution of the automotive supply network confirmed the small world effect. However, this did not show up in the aerospace industry when studied in [2]. The reason for this difference might be due to the fact that aerospace production is a bespoke and low volume business, leading to highly specialized suppliers and hence a low degree of clustering within the network. On the other hand, the automotive industry is standardized and high volume, leading to common supplier switch-over and multi-sourcing and hence a high degree of clustering within the network. Currently the degree distribution has not been generalized for supply chain networks of different industries and this remains an interesting area for future research [3].

At the node level, several authors have used the metrics such as degree centrality, betweenness centrality, authority and hub centralities, etc to identify firms occupying special positions in the supply chain network. For example, in [4], the Katz centrality was proposed to measure a supplier's risk of spreading disruption while the authority and hub centralities were applied to measure the risk of a link failure.

For the mesoscale characterization of the supply chain networks, the work has generally focused in the detection and analysis of communities within the network. For example, the authors in [1] detected different communities in the Toyota network using a community detection algorithm for directed graphs [5]. Analysis of the communities showed that their structures differed markedly depending on whether a particular community produced parts specific to the automotive industry or parts used more widely across different industries, e.g. electronics. For communities producing electronic parts which are broadly used, the network structure is decentralized with a high clustering coefficient. The work in [1] is most similar to the approach that we will take in the paper with respect to the analysis of the supply chain network.

Finally, in related work [6], although it did not consider the topology of a supply chain network, it did provide inspiration for us to use historical stock price changes as an empirical measure of the accuracy of the network models. The authors in [6] showed that the upgrade (downgrade) of a significant customer will lead to a lagged positive (negative) return in the stock price of the supplier. The delayed response is posited to be due to the attention constraints of investors to all the different types of links between corporations. Indeed the authors showed that this effect can lead to market outperformance if embedded in a trading strategy which goes long (short) the supplier's stock when the stock of large customer rises (falls).

3 Methodolgy

3.1 Dataset Collection

The data for this paper is mostly obtained from Bloomberg. Unlike other types of financial data, supplier-customer data are usually incomplete because corporations have few regulatory obligations to disclose and are unlikely to do so voluntarily as supplier-customer relationships can be seen as trade secrets. For example, in the US, the SEC's Regulation S-K states only that corporations must disclose customers that are greater than 10% of revenues on an annual basis. Since around 2011, Bloomberg started to make available to its users supply chain data aggregated from Regulation S-K disclosures as well as other sources, e.g. conference call transcripts, capital markets presentations and company press releases. In Appendix, we show the current largest customers of Intel Corp sorted by their contribution to Intel Corp's revenues (e.g. Dell Technologies Inc is the largest customer by revenue contribution at 16%).

For this paper, we only considered corporations (either as supplier or customer) which are included within the Russell 3000 index. This index is composed of the 3000 largest US listed companies as determined by market capitalization, approximately 98% of the investable US equity market. We imposed this constraint as (i) the supply chain data for US firms is more reliable due to regulatory disclosure requirements and (ii) historical stock price data is only available for listed firms. However, we realize that this constraint is a trade-off as large private companies such as Dell Technologies Inc will be excluded as well as companies domiciled outside of the US, e.g. Lenovo Group Ltd. Hence, the top two customers in Fig. 2 Appendix are excluded from our network. One potential avenue for future research is to include corporations listed outside of the US in the network to reflect the increasingly global nature of supply chain.

For each firm in the Russell 3000 index, we extract the identities of all of its customers from the Bloomberg database. Then for each customer, we determine whether it is in the Russell 3000 index and exclude it from the network if it is not. Our customer data cover the period between 2013 and 2018. We also extract from Bloomberg the historical stock prices for every supplier and customer within our network for the whole time period on a monthly frequency.

3.2 Algorithms

In the main analysis of this paper, we frame the work as a prediction problem. Specifically, for each pair of nodes in the network, we are trying to predict the correlation between the returns of their stock prices using the structure of the network as features.

3.2.1 Shortest Path Distance and Node Centrality

We start with the simpler measures of the network as features. The shortest path distance is the most intuitive of these as we would expect that companies which are neighbours to have much higher correlated stock prices than pairs which are several hops away or not connected at all. In economic terms, this is because when the customer of a firm does well (or badly) then the firm itself will do well (or badly).

There are many measures of centrality for nodes in a network. These are used to rank the importance of the nodes because we would expect that a large firm such as Walmart Inc, which is central to a large interconnected network of many suppliers, would move the stock price of those suppliers more than smaller firms. We choose PageRank [6] to test this as it is one of the measures that has shown impressive empirical results in other networks.

For any node j, the PageRank value r_j is defined iteratively as: $r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1-\beta)\frac{1}{n}$

3.2.2 Motif Detection and Motif Spectral Clustering

To look at structures which are high level than single nodes, we use motif detection [7] and clustering [8]. In motif detection, we count the frequency of non-isomorphic direct subgraphs of a certain size in our network and compare that frequency distribution with a null model. Specifically, we count all 3-subgraphs in our network and compare that with the configuration model which generates random networks with the same degrees sequence as our original network. The configuration model does this by taking our original network, picking pairs of edges at random and then switching the end points. This repeated many times, which in our case was around 10,000.

Comparing the frequency of different motif types in our network allows us to find motifs which are over-represented in our network and hence might be an important aspect of supply networks in general. We can then cluster the network using those over-represented motifs as targets in the spectral clustering algorithm [8]. This is an algorithm which constrains clustering based on keeping as many instances of a particular motif intact as possible within each community. More precisely, given a motif M, the algorithm aims to find a cluster (defined by a set of nodes S) that minimizes the "Motif Conductance":

$$\varphi_M(S) = \frac{cut_M(S, \bar{S})}{min[vol_M(S), vol_M(\bar{S})]}$$

Finding the set S with the minimal motif conductance is NP-hard. However, in [8], the researchers showed that it is possible to use some of the machinery from spectral clustering to produce a similar algorithm which will obtain near optimal motif conductance. The algorithm is as follows:

- Step 1: Given a graph G and motif M, form a new motif adjacency matrix W_M whose entries (i, j) are the co-occurrence counts of nodes i and j in the motif M: $(W_M)_{ij}$ = number of instances of M that contain nodes i and j
- Step 2: Apply spectral clustering on W_M to compute the spectral ordering z of the nodes
- Step 3: Find the prefix set of z with the smallest motif conductance, i.e. if $S_r = \{z_1, \ldots\ldots, z_r\}$:

$$S := \arg min_r \varphi_M(S_r)$$

Once we have the clusterings, we generate a new feature for each pair of nodes which is 1

3.2.3 Node2Vec embeddings

Finally node2vec embeddings [9] can also be used as features in our prediction. The algorithm to generate node2vec embeddings is based on the random walk approach which proposes that, in general, the dot product of the embeddings of a pair of nodes (*u, v*) should be close to the probability that *u* and *v* co-occur on a random walk. This leads to the minimization of the objective function, which in practice is minimised using negative sampling:

$$L = \sum_{u \in V} \sum_{v \in N_R(u)} -log\left(\frac{exp(z_u^T z_v)}{\sum_{n \in V} exp(z_u^T z_v)}\right)$$

The node2vec algorithm also introduces the concept of biased random walks so that the embeddings could interpolate between BFS and DFS exploration of the network. This is done by adding a return parameter p and an in-out parameter q. If the current node of the random walk is w and the previous node was s, then the probability of (i) moving to back to s is $1/p$, (ii) moving to any node which is the same distance from s as w is 1 and (iii) moving to any node which is farther away from s than w is $1/q$.

To combine the embeddings from two nodes as a feature, we concatenate or sum the embeddings from the individual nodes.

3.2.4 Learning Algorithm

Once we have all the features generated above, we train a ridge regression model as a learning algorithm to predict the correlation of stock returns between two nodes. Concretely, the ground truths vector *Y* contains the correlation of returns for all possible node pairs (*u*, *v*). The feature array *X* contains the features generated from each pair (*u*, *v*), e.g. shortest path length between *u* and *v*, and sum of embeddings of *u* and *v*, etc. We divide the data into a training set (around 70% of the total data) then use the remaining data as the test set. Each experiment will consist of training a model on the training set using only a subset of the features, e.g. PageRank only, and then that model will be evaluated on the test set. The experiments overall tell us whether using more sophisticated features generated from the structure of the network leads to better predictions.

Ridge regression addresses some of the problems of ordinary least squares regression by imposing a penalty on the size of the coefficients. Hence, the penalized residual sum of squares that need to be minimized becomes:

$$min_w \|Xw - y\|^2 + \alpha \|w\|^2$$

The score of a ridge regression model is given by the coefficient of determination R^2 of the prediction. The best possible score is 1.0 and a constant model (not taking into account any features) would get a score of 0.0.

4 Results

4.1 Summary Statistic of the Network

We build the network using every corporation within the Russell 3000 index as a node. We then create a directed edge from one firm to another if the former (supplier) sells goods to the latter (customer). This network has 3007 nodes and 4791 edges. (Despite its name, the Russell 3000 actually has 3007 constituent members.) To visualize part of the network, we show in Appendix the customer subgraph of Intel Corp. It makes sense that well-known hardware companies such as AAPL and HPE should be large customers of Intel Corp, and we note once again that Dell Technologies Inc is excluded from the network as it is not a listed company.

Looking at the network level, Fig. 2 shows that the network exhibits a power-law degree distribution of node out-degrees, i.e. $P(k) = ak^{-y}$ for constants a and y. Hence it is scale-free with $y = 1.30$. Fig. 3 shows that this is also true when considering node in-degrees, with $y = 1.15$ in this case.

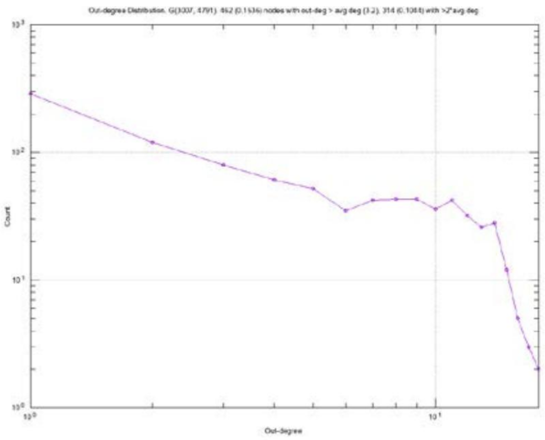


Figure 2: Degree distribution of Node Out-degree (log scales)

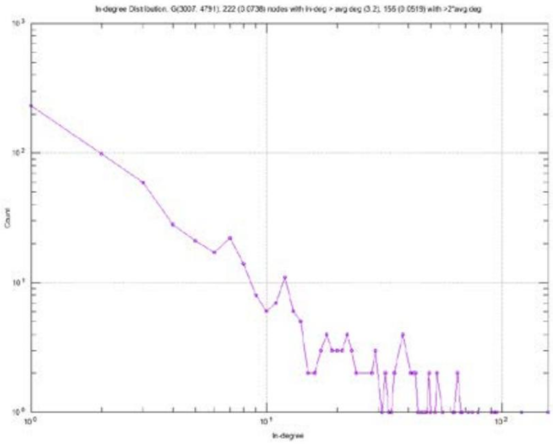


Figure 3: Degree distribution of Node In-degree (log scales)

Fig. 4 shows the average clustering coefficient as a function of node degree. Here we see that the clustering coefficient decreases with increasing node degree, which again is indicative of a scale-free network.

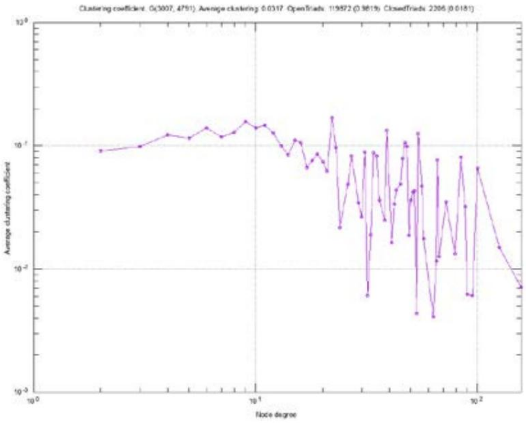


Figure 4: Average clustering coeff vs Node degree

We further find that the largest strongly connected component consists of 91 nodes and 294 edges, and the largest weakly connected component consists of 1178 nodes and 4767 edges. Since the largest weakly connected component only consists of 40% of the total number of nodes but 99% of the total number of edges, many of the nodes in the network are not connected to other nodes. We note that this could be due to the incompleteness of our data, i.e. a supplier-customer relationship might not be captured in the

Bloomberg database if they cannot be found in public sources. We keep this in mind when performing other analyses of the network, e.g. community detection.

To create the correlation matrix for the stock price returns, we calculated the monthly returns for each stock over 2013 to 2018. The correlation was then calculated for each of the stock's return time series against every other stock. This resulted in a 3007 by 3007 correlation matrix.

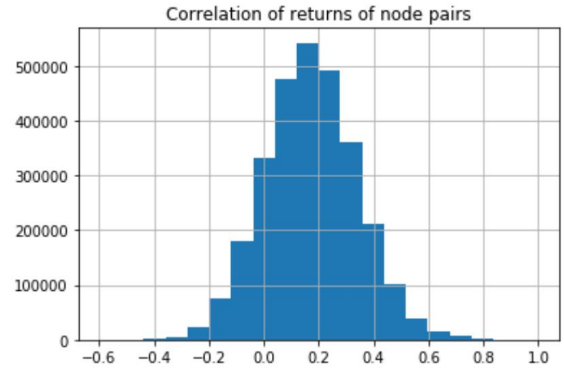


Figure 5

Figure 5 is a histogram of the correlations over all pairs of nodes. It shows that the distribution of correlations is well balanced with mean around 0.17%.

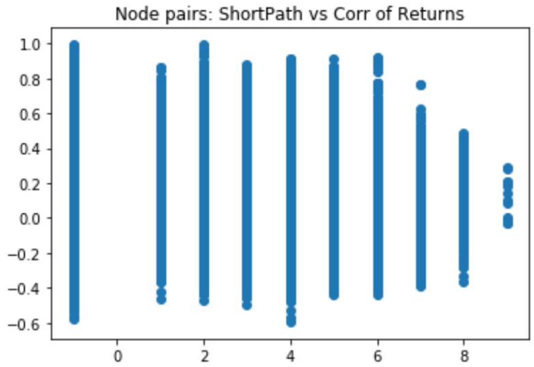


Figure 6

Figure 6 is a plot of the distribution of the correlations for different shortest path lengths. It shows (i) for the smaller path lengths, the range of correlations can be very large so there's no easy solution based on shortest path length alone and (ii) for longer path lengths, the correlations are more tightly centred around 0, which is to be expected.

4.2 PageRank

Using PageRank [6] as a measure of the importance of nodes, we find that the five corporations with the highest PageRanks are:

1) Walmart Inc
2) The Boeing Co
3) Lockheed Martin Corp
4) Intel Corp
5) Tech Data Corp

It makes sense that such large and complex organizations should have the highest PageRanks.

As discussed in the methodology section, we then use the PageRanks of a pair of nodes to try and predict the correlation of stock returns for those nodes. The ridge regression model was trained on the

PageRanks of all the pairs of nodes in the training data as well the shortest path length between each pair. Using the model on the test data, we found the R2 was 0.00023. Hence the model does not perform better than a constant model.

4.3 Motifs

We next examine the network using motif detection [7] and compare the frequency of non-isomorphic directed 3-subgraphs in our network with a null model. In this case, the null model is created using the configuration model which generates random networks with the same degree sequence as our original network. Fig. 7 shows the z-scores obtained from our network for each motif type.

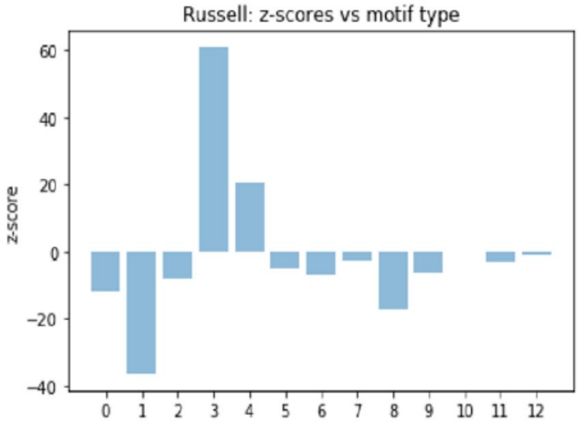


Figure 7

The spike represents a motif where two suppliers selling to one manufacturer (i.e. two nodes both pointing to a third node with no other edges) so it makes sense that our network would have overabundance of this motif versus a random network.

Hence it made sense to target similar motifs using the motif-based spectral clustering algorithm. We performed motif-based spectral clustering on the network using M5, M6, M8, M9, M10 (where M10 is the over-represented one).

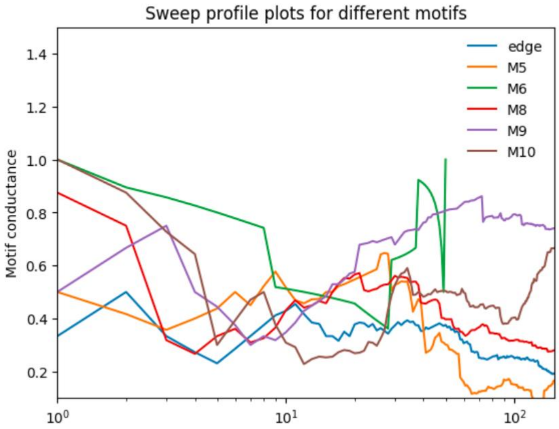


Figure 8

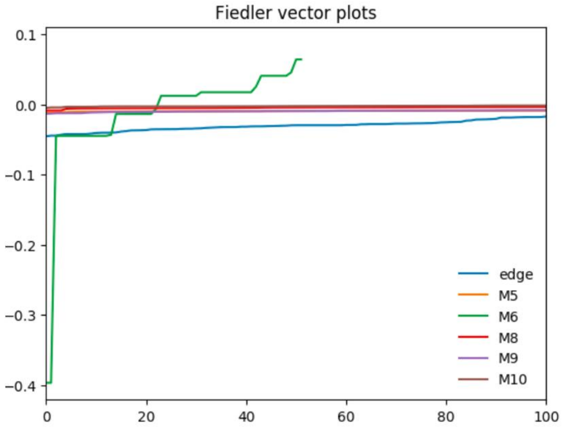
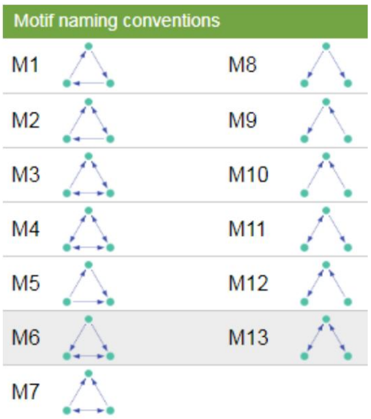


Figure 9



Motif naming conventions	
M1	M8
M2	M9
M3	M10
M4	M11
M5	M12
M6	M13
M7	

Figure 10

Plotting the sweep profile for each motif in Figure 8 showed that M10 does not cluster the network better than the other motifs its over-representation. This can also be seen in the plot of the Fiedler vectors for each of the motif solutions in Figure 9. Only the Fiedler vector for M6 has well distinguished plateaus between the elements of the Fiedler vector. But unfortunately the M6 motif is not well represented in the network hence conductance is not well minimized when targeting this motif.

Using the ridge regression test, we find that again the model does not perform well. In fact, adding the motif-clustering feature to PageRank slightly decreases R2 score to 0.00021.

4.4 Node2vec

Node2vec embeddings were generated from the network using different sets of hyperparameters, e.g. (*p*, *q*) = (1, 5), (1, 1), (5, 1), embedding dimensions = 64, 128, etc. To check the embeddings qualitatively, we used k-means clustering algorithm to separate the nodes into 10 clusters. The network was then drawn with those clusters in place. The node2vec clusterings did a good job of identifying the communities of dense edges as shown in the plot.

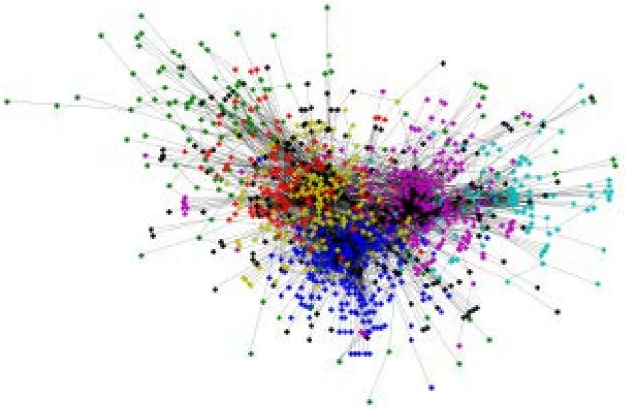


Figure 11: Clustering of supply network using node2vec: p1, q1

Node2vec also showed better performance than the other methods of generating features when used to train the ridge regression model.

p	q	Embedding dimension	Walk length	R^2
1	1	128	80	0.565194
1	5	128	80	0.520322
5	1	128	80	0.552524
1	1	64	80	0.564501
1	1	128	160	0.521488

It is not clear why the node2vec embeddings do much better than the other types of features, one possible reason is due to the larger number of variables. Changes in the hyperparameters do not make a large difference to the performance.

5 Conclusions

We generated a range of features from the network to train a regression model for predicting the correlation of stock price returns between firms in the supply network. These included PageRank, Motif clusters and Node2Vec. Features from PageRank and Motif clusters do not have much predictive power at all. Node2vec performs significantly better and further work can be done to investigate the source of the outperformance. In addition, it would be interesting to investigate whether node2vec can be used to predict other features of the supply network, e.g. evolution of new supplier-customer relationships.

6 References

[1] Kito, T., Brintrup, A., New, S., and Reed-Tsochas, F., The Structure of the Toyota Supply Network: An Empirical Analysis (Sa€ıd Business School, WP, 2014), p. 3.

[2] Brintrup, A., Barros, J., and Tiwari, A., “The nested structure of emergent supply networks,” IEEE Syst. J. PP, 1 (2015).

[3] Brintrup, A., Ledwoch, A., "Supply network science: Emergence of a new perspective on a classical field"

[4] Ledwoch, A., Brintrup, A., Mehnen, J., and Tiwari, A., “Systemic risk assessment in complex supply networks,” IEEE Syst. J. PP, 1 (2016).

[5] Leicht, E. A. and Newman, M. E., “Community structure in directed networks,” Phys. Rev. Lett. 100(11), 118703 (2008).

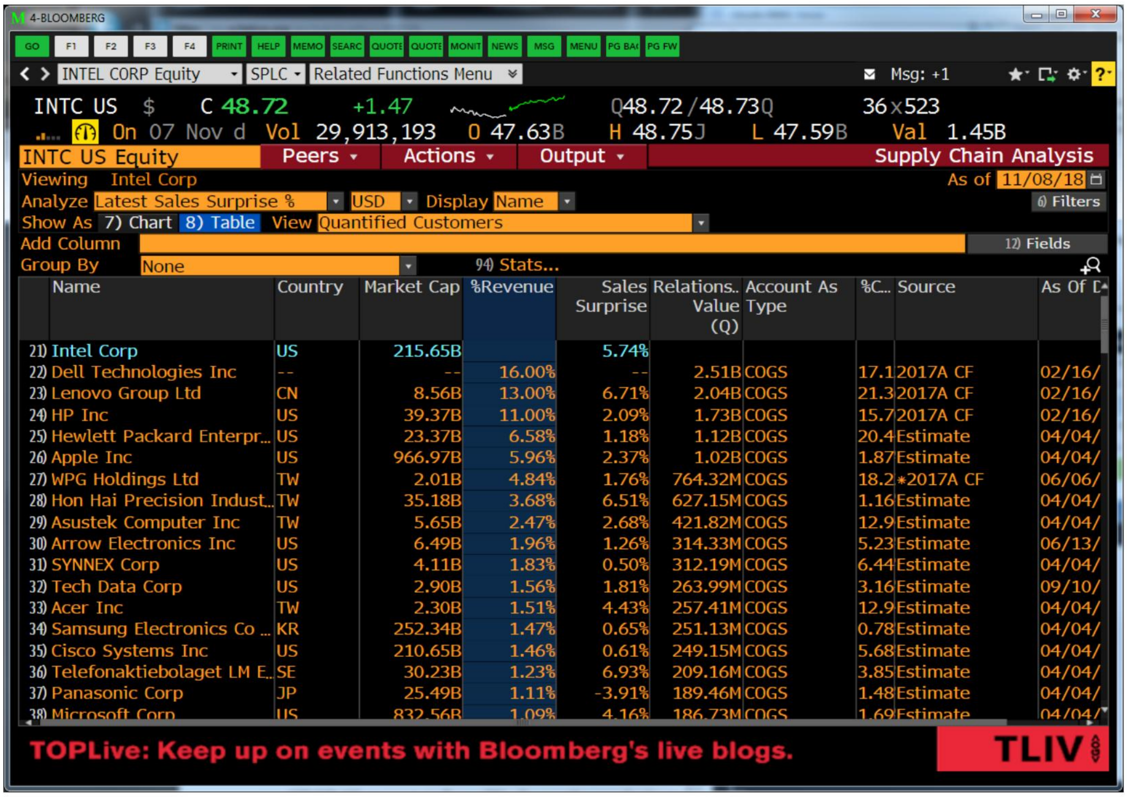
[6] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation Ranking: Bringing Order to the Web

[7] Milo, R., et al, Network Motifs: Simple Building Blocks of Complex Networks

[8] Austin R. Benson, David F. Gleich, and Jure Leskovec., Higher-order Organization of Complex Networks.

[9] Grover et al, node2vec: Scalable Feature Learning for Networks. KDD. (2016)

7 Appendix



INTEL CORP Equity | SPLC | Related Functions Menu

INTC US $ C 48.72 +1.47 Q48.72/48.73Q 36x523

On 07 Nov d Vol 29,913,193 O 47.63B H 48.75J L 47.59B Val 1.45B

INTC US Equity

Viewing Intel Corp As of 11/08/18

Analyze Latest Sales Surprise % USD Display Name 6) Filters

Show As 7) Chart 8) Table View Quantified Customers

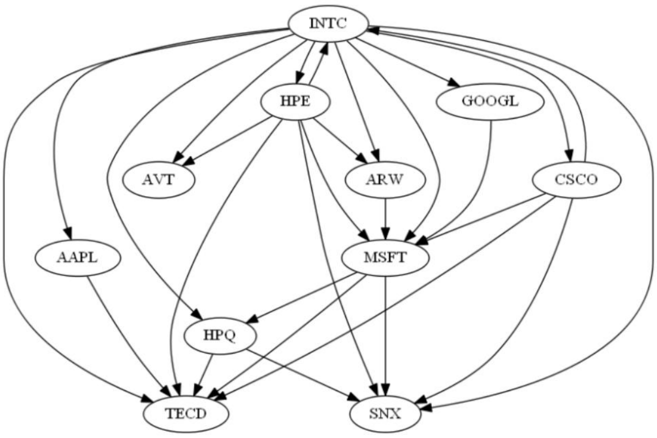
Add Column 12) Fields

Group By None 94) Stats...

	Name	Country	Market Cap	%Revenue	Sales Surprise	Relations.. Value (Q)	Account As Type	%C...	Source	As Of D
21)	Intel Corp	US	215.65B		5.74%					
22)	Dell Technologies Inc	--	--	16.00%	--	2.51B	COGS	17.1	2017A CF	02/16/
23)	Lenovo Group Ltd	CN	8.56B	13.00%	6.71%	2.04B	COGS	21.3	2017A CF	02/16/
24)	HP Inc	US	39.37B	11.00%	2.09%	1.73B	COGS	15.7	2017A CF	02/16/
25)	Hewlett Packard Enterpr...	US	23.37B	6.58%	1.18%	1.12B	COGS	20.4	Estimate	04/04/
26)	Apple Inc	US	966.97B	5.96%	2.37%	1.02B	COGS	1.87	Estimate	04/04/
27)	WPG Holdings Ltd	TW	2.01B	4.84%	1.76%	764.32M	COGS	18.2	*2017A CF	06/06/
28)	Hon Hai Precision Indust...	TW	35.18B	3.68%	6.51%	627.15M	COGS	1.16	Estimate	04/04/
29)	Asustek Computer Inc	TW	5.65B	2.47%	2.68%	421.82M	COGS	12.9	Estimate	04/04/
30)	Arrow Electronics Inc	US	6.49B	1.96%	1.26%	314.33M	COGS	5.23	Estimate	06/13/
31)	SYNNEX Corp	US	4.11B	1.83%	0.50%	312.19M	COGS	6.44	Estimate	04/04/
32)	Tech Data Corp	US	2.90B	1.56%	1.81%	263.99M	COGS	3.16	Estimate	09/10/
33)	Acer Inc	TW	2.30B	1.51%	4.43%	257.41M	COGS	12.9	Estimate	04/04/
34)	Samsung Electronics Co ...	KR	252.34B	1.47%	0.65%	251.13M	COGS	0.78	Estimate	04/04/
35)	Cisco Systems Inc	US	210.65B	1.46%	0.61%	249.15M	COGS	5.68	Estimate	04/04/
36)	Telefonaktiebolaget LM E...	SE	30.23B	1.23%	6.93%	209.16M	COGS	3.85	Estimate	04/04/
37)	Panasonic Corp	JP	25.49B	1.11%	-3.91%	189.46M	COGS	1.48	Estimate	04/04/
38)	Microsoft Corp	US	832.56B	1.09%	4.16%	186.73M	COGS	1.69	Estimate	04/04/

TOPLive: Keep up on events with Bloomberg's live blogs. TLIV

Largest customer of Intel Corp by revenue contribution (source: Bloomberg)



Customer subgraph of Intel Corp