Analyzing the Corporate Ecosystem

Salik Syed
Stanford University
Stanford, CA
ssyed@stanford.edu

Nikil Viswanathan
Stanford University
Stanford, CA
nikil@stanford.edu

Tony Wu
Stanford University
Stanford, CA
tonywu@cs.stanford.edu

ABSTRACT

Our project examines the implicit interactions of the corporate ecosystem constructed with data gathered from stock prices across the last decade. The correlation of stock price residuals is used to construct a corporate information network. We found that residual of stock price among each company's tradable security in the S&P 500 followed a roughly normal distribution, but that the distribution of correlations was different with each respective sector (technology, energy, health, etc.).

We found that several network metrics (clustering coefficient), evaluated on the graph over different time periods coincide with major stock market trends. By implementing community detection techniques, we discovered interesting intra- and inter-sector relationships between clusters of specific companies. We have generated an interesting network capturing subtle interactions between the S&P 500 companies that can be used for further inspection and modeling of the US economic powerhouse companies.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks; H.3.4 [**Systems and Software**]: Information Networks

General Terms

Algorithms, Information Network Analysis

Keywords

Stock price residual, correlation coefficient, econometric analysis, influence model, influence maximization, outbreak detection

1. INTRODUCTION

Needless to say, there has been a great amount of study has been devoted to the stock market and securities trading. Decades of analysis have been put into the understanding of what influences the share prices of all types of securities. However, there has been very little research that models the entire stock market as a network of securities, while applying social and information network analysis to such a network. We aimed to pioneer a new form of macroeconomic analysis by building networks of corporations based on relationships of their stock prices and competitive nature.

2. DATASET

2.1 S&P 500

Our dataset comes from the stock prices of the publicly traded securities that compose the S&P 500 index gathered via the Yahoo Finance API, spanning from 1998 to 2010. Since there's such a vast amount of data, only the opening and closing prices for each traded day are recorded and used in our analysis.

The S&P 500 opening and closing prices are extracted and the correlation is computed between each pair of securities. These serve as edge weights in our network analysis.

2.2 Corporate Graph: Standard vs. Residuals Approach

We first took a straightforward approach and constructed our corporate graph by simply observing the opening-day pair-wise correlations of each of the S&P 500 securities. We add an edge between securities a and b whose weight is the Pearson correlation coefficient computed for the said securities. The resulting correlation graph was not very effective as it captures the entire macroeconomic trend of that time period, and as a result provides a great deal of noise and results in the vast majority of security pairs having an extremely positive correlation. See figure 1.

In our second model, we let the residual of a stock denote the difference between the price and the price that is attained by computing a best-fit line through the data. We use the correlation of the residual instead of direct correlation of price. Such a statistic ignores large macroeconomic trends, which would otherwise add noise to our data. Observe that the edge weights are less skewed by the macroeconomic trends that affect all stocks.

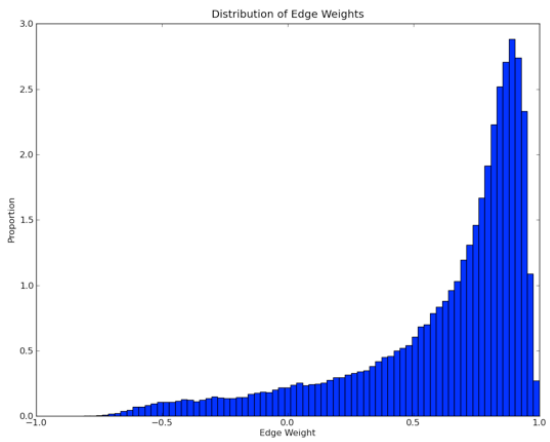

Figure 1. Distribution of Price Correlations, 2009-2010 (Straightforward Approach)

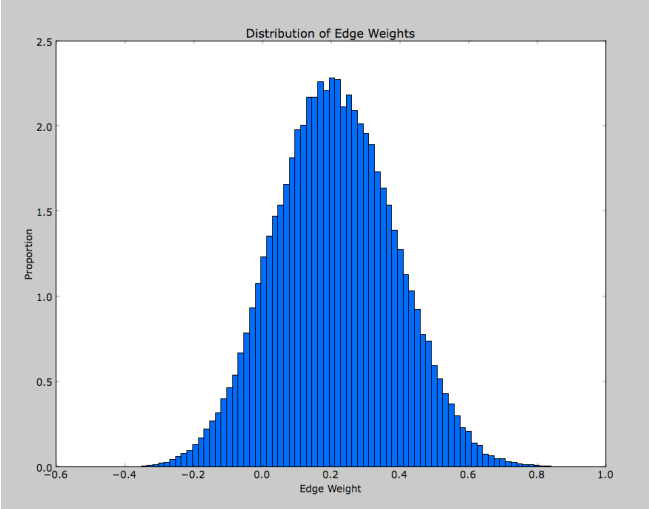

Figure 2. Distribution of Price Residuals, 2009-2010

As a result, the skew of the data disappears when taking the residual and we have a normally distributed set of edge weights with a mean of 0.21.

Here are samples of some of the nearest neighbors generated using correlation of price residual:

Table 1. Sampling of Nearest Neighbors using Correlation Residuals.

Company	**Nearby Nodes**
IBM (International Business Machines)	ADE (Adobe), CSC (Cisco), APPL (Appl)
JWN (Nordstorm)	RL (Polo Ralph Lauren), SBUX (Starbucks)
MRK (Merck)	JNJ (Johnson & Johnson), MDT (Medtronic)
GS (Goldman Sachs)	JPM (JPMorgan Chase) ,MS (Morgan Stanley)
KFT (Kraft Foods)	K (Kellog), SLE (Sara Lee)
CVX (Chevron)	COP, (ConocoPhillips),XOM (Exxon Mobil), SE(Spectra Energy)
Raytheon	NOC (Northrop Grumann), GD (General Dynamics), (Lockheed Martin)

3. NETWORK ANALYSIS

Network analysis was conducted in order to find trends with respect to the real-life stock market, and also discover interesting reads. Based on the analysis conducted, we were able to draw conclusions about the nature of stock price interaction as well as devise a model for how a network of publicly tradable securities behaves.

3.1 Distribution of Edges in Network

Using the correlation graph generated from price residuals, as seen in figure 1, the distribution of edge weight (tie strength) follows a normal distribution, centered around 0.21 with a standard deviation of 0.23. Preserving only ties that are greater than a standard deviation above the mean tie strength gives us a non-complete graph with 16% of remaining edges. The remaining edges follows that of an approximate power law, with alpha = 0.39, observed in Figure 3.

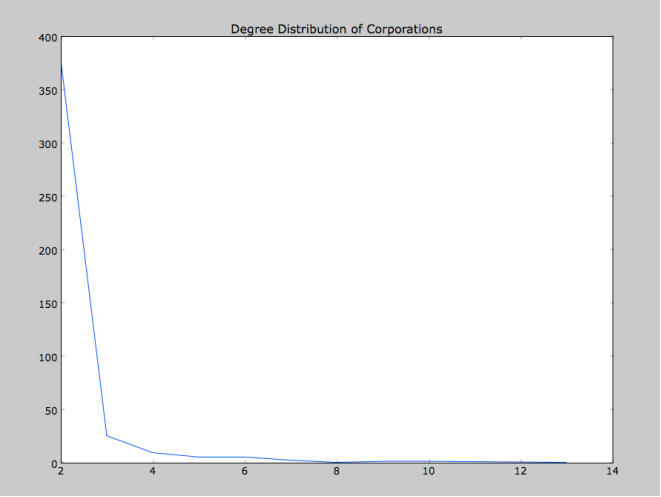

Figure 3. Degree Distribution follows Power Law

3.2 Power Law and Influence Model

In order to justify the given power law distribution, we formulated an *influence model*. In our model of influence, we theorize that as a corporation grows, it increases its network of influence. In turn, more investors look towards that company's

finances and stock price as an indicator of how the market is doing, and how the stocks within that investor's portfolio are doing. As a result, that stock becomes more and more correlated with those many stocks. It's widely known that investors make decisions only based on a subset of the publicly available data that they find to be relevant and that influences a subset of those tradable securities [4]. For each tradable security, the likelihood that an investor will begin to make investment decisions based on that stock is proportional to the number of stocks it is already correlated and also roughly proportional to the number of investors who are already observing that stock and making portfolio decisions based on that stock. Since power laws are a result of "rich get richer" situations where a node's probability of connecting to a given node is proportional to the degree of that node [3], our hypothesized *Influence Model* for publicly tradable corporate securities effectively explains why power laws exist in the stock correlation network.

3.3 Network Analysis and Stock Market Trends

Using further network analysis stock market trends were discovered to coincide with interesting fluctuations in the market. For example, from 2000 – 2002 there was a slump in stock market culminating in the post-Sept. 11 stock dip along with the stock market crash in March of 2002. Spikes in **clustering coefficient** of the network coincided strongly with bouts of stock market decline, particularly near the crashes.

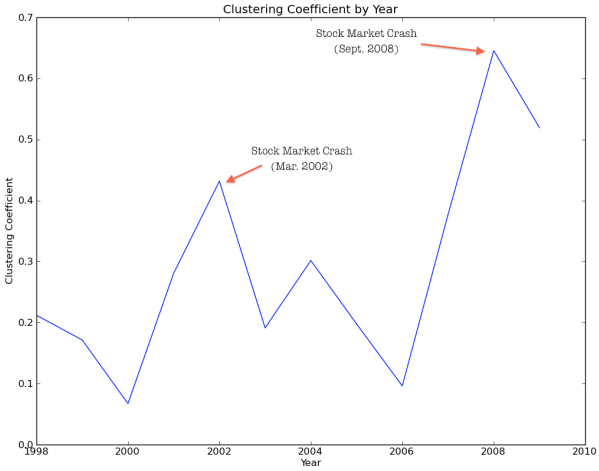

Figure 4. Clustering Coefficient Over Past Decade

Rapid, large-scale increases in clustering coefficient seems to be indicative of an impending stock market crashes without very much noise. The clustering coefficient already began to rise rapidly almost two years before each of the stock market crashes (starting in 2000 and 2006) and did so without much noise.

One plausible non-information-network-related explanation for the relationship between increases in the network's clustering coefficient and stock price (or return on investment, ROI) is that during an economic downturn, all of the stocks go down simultaneously, which results in a dramatic increase in the correlation of all the stocks, which by extension, increases the network's clustering coefficient. However, this is debunked by the fact that there is very little correlation between the *average correlation coefficient* and ROI. Over the past decade the correlation coefficient **r** between the *average correlation coefficient* of the S&P 500 securities and *ROI* is a mere -0.286.

However, using network and graph analysis, we can find much more insightful information. We find that the correlation coefficient **r** between the *average correlation coefficient* of the S&P 500 securities and *ROI* is an extremely strong -0.701. Observe Figure 5 and Figure 6.

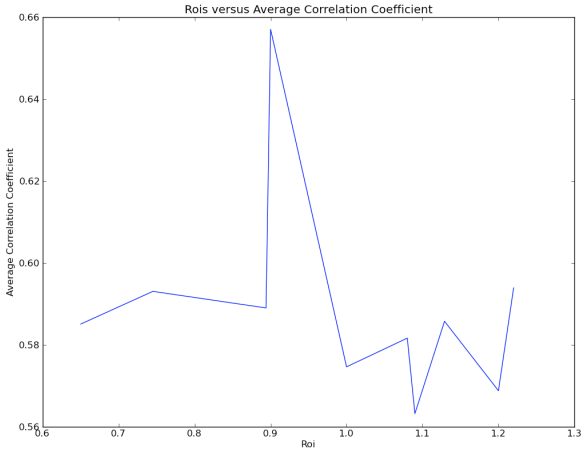

Figure 5. S&P 500 ROI vs. Traditional *Stock Correlation Coefficient*: The correlation *between average correlation coefficient* and ROI is a mere *-0.286*.

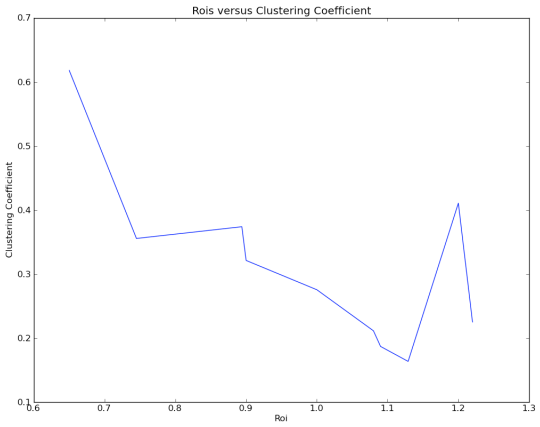

Figure 6. Stock ROI vs. Network's *Clustering Coefficient*: The correlation *between average clustering coefficient* and ROI is a strong *-0.701*.

This demonstrates that using network analysis can improve upon many areas where traditional statistical econometrics fall short. Figure 5 and Figure 6 demonstrate that the statistics derived from a network-based model of the stock market can be less susceptible to noise and act as a stronger signal for economic ROI than traditional means of analyzing data. This is one area where we demonstrate that the application of network analysis can have profound impacts in macroeconomic theory.

4. CLUSTERING ANALYSIS AND COMMUNITY DETECTION

In order to see if we could divide the nodes into actual discernable sectors, we looked towards clustering as a means of doing this. To segment the nodes in our graph we performed two different types of clustering: Agglomerative Hierarchical and MCL. We found that MCL clustering in general resulted in better-defined clusters and were more representative of the types of the securities and also provided interesting information about the relationships of different sectors.

4.1 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering consists of repeatedly merging clusters via dendogram [2] (finding lowest costing edge and merging the two corresponding clusters together). Agglomerative Hierarchical Clustering can also use single linking and complete linking - of the two complete linking was more effective. Single linking results in one large cluster because the larger a cluster is, the more likely the closest edge resides within that cluster and so the clustering results in more and more nodes binding to the larger clusters. We found complete linking to be much more effective. Sample clusters generated with Agglomerative Hierarchical Clustering:

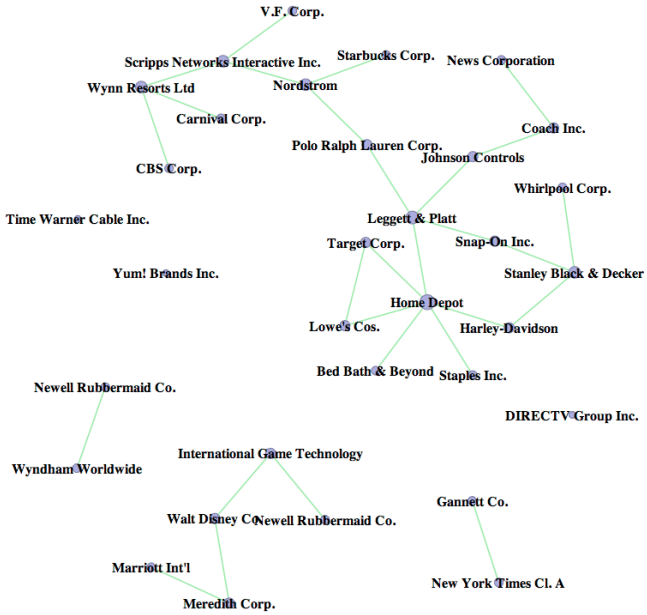

Figure 5. Consumer Discretionary Clusters

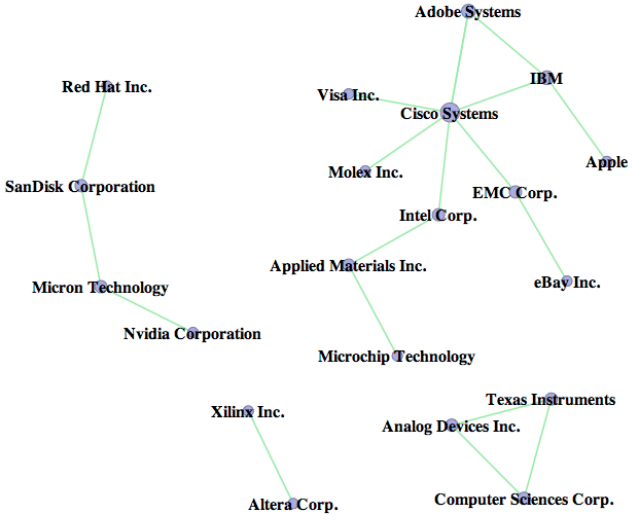

Figure 6. Technology Clusters

Using Agglomerative clustering with complete linkage, we were able to get well-defined clusters. For example, in the Consumer Discretionary sector Home Depot, Target, Lowe's reside in one community. Another cluster is composed of New York Times and Gannet Co. (parent company of USA Today).

Similarly, in the tech sector, we find that many of the high tech companies are clustered together (Adobe, IBM, Cisco, etc.). We also see that Texas Instruments, Analog Devices, and Computer Sciences Corp are clustered (they all make scientific devices). Altera and Xilinx are clustered together (they make FPGA and CPLD solutions).

The tendency of companies in the same sector and do similar products to cluster together can be attributed partly to homiphily,

but also to common investment behavior, that was discussed in section 3.2 and also manifests itself in the power law of the degree distribution.

4.2 Community Detection with Markov Clustering

We improved our clustering by using Markov clustering with the MCL algorithm [1], designed specifically for community detection, which in theory apply better to networks than the hierarchical clustering techniques. With agglomerative clustering, even with lower thresholds, even with lower thresholds for the correlation cutoff, we kept on seeing the rise of one massive central cluster and fragmented nodes surrounding it. This is likely attributed in part to preferential attachment to the largest cluster, along with the existence of small "whiskers" in the graph that don't cluster well with the large "core" of the network [5]. In an attempt to get a more even clustering, we used a different clustering algorithm – the Markov Clustering Algorithm (MCL). The algorithm uses random walks through the graph to discover communities [11]. We notice that with the new clustering algorithm, we get a clustering breakdown that has a nice spread of cluster sizes (Table 1).

Table 1. Agglomerative Hierarchical Clustering Statistics

Cluster Size	Number of Clusters
141	1
2	4
1	285

Table 2. Markov (MCL) Clustering Statistics

Cluster Size	Number of Clusters
194	1
60	1
23	1
8	1
7	1
5	3
4	7
3	7
2	11
1	56

We also observed the relationship between the sectors and the clusters generated from MCL:

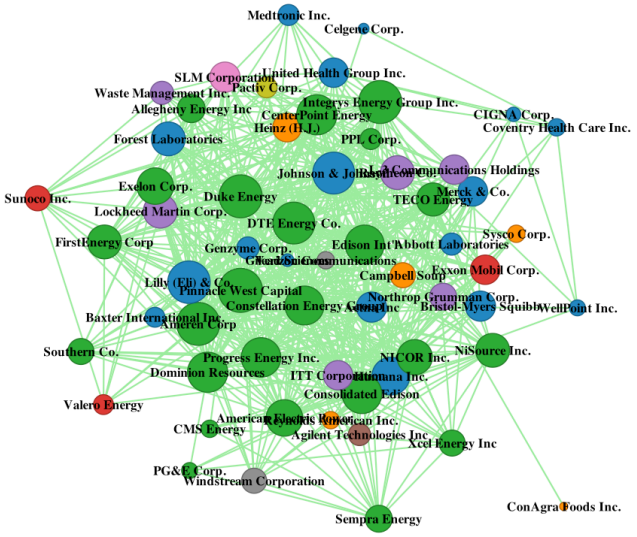

Figure 7. Sample Cluster (1)

Table 3. Legend for Figure 7.

	Utilities
	Health
	Energy
	Consumer Discretionary
	Communications
	Technology

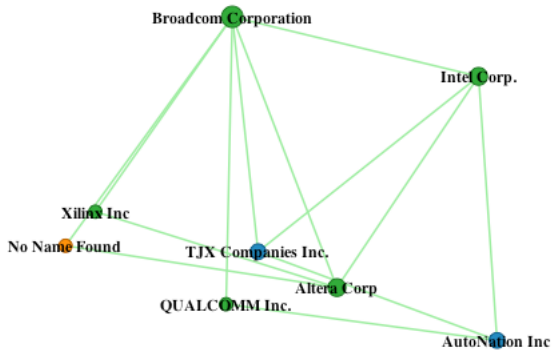

Figure 8. Sample Cluster (2)

Table 4. Legend for Figure 7.

	Tech
	Consumer Discretionary

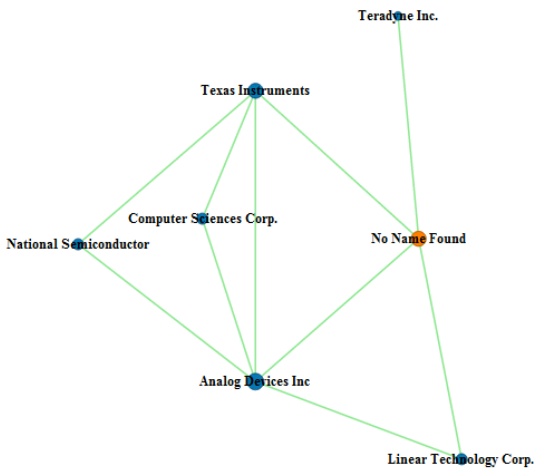

Figure 9. Cluster of Semiconductors

This cluster found the relationships between the large semiconductor and chip manufacturers and the companies that provide services for these special components. The orange circle is Maxim corporation, another semiconductor manufacturer and is actually in the same sector as the other companies but is colored differently since the most recent S&P 500 stock company name list did not include it.

One interesting piece of insight is that utilities and health companies seemed to form in the same clusters, while tech and consumer discretionary companies also tended to cluster together into a different community.

Since many traditional clustering algorithms are not robust against small outlying whiskers [8, 10], we believed that applying MCL as a specialized algorithm for community detection is prudent and this hypothesis was backed up empirically by the clusters that we generated.

5. ONLINE CLUSTER VISUALIATION

For our full set of online cluster visualizations, observe http://stanford.edu/~nikil/cs224w/trunk/web/stockClusters.html

6. CONCLUSIONS

Our work in this area has produced several meaningful insights and has the potential for wide-ranging practical applications in the field of econometrics. By utilizing clustering analysis we've demonstrated that modeling a stock market as an information network is a viable model for providing accurate and interesting results. Graph-clustering techniques provide intuitively sensible clusters based on corporate similarities. This in turn demonstrates the underlying viability of our Influence Model and the stock correlation network as a representation for the interaction of real world securities.

7. FUTURE

7.1 Outbreak Detection and Influence Maximization

We wish to further explore the area of stock market trends via **outbreak detection** and **influence maximization** analysis. Now that we've established that modeling the stock exchange as an intricate information network, we wish to apply analysis of influence maximization sets and outbreak detection to see if they apply to the model at hand. If the outbreak that is the financial crisis of 2001-2002 and 2008-2009 follows our model of influence, then we should be able to detect a handful of stocks that serve as the initial influence set.

8. ACKNOWLEDGMENTS

Our thanks to Professor Jure Leskovec along with the course assistants Nadine, Sudarshan, Sonali and Jennifer for an enlightening and exciting quarter in CS224W.

9. REFERENCES

[1] Dongen, S. 2000. *Graph Clustering by Flow Simulation*. University of Utrecht. DOI= http://igitur-archive.library.uu.nl/dissertations/1895620/full.pdf.

[2] Koga, H., Toshinori W., and Ishibashi, T. 2007. *Fast agglomerative hierarchical clustering algorithm using Locality Sensitive Hashing*. University of Electro-Communications. DOI=http://portal.acm.org/citation.cfm?id=1266811

[3] Clauset, A., Shalizi, C., Newman, J.E. 2007. Power-law distributions in empirical data. In SIAM Review 51, 661-703 (2009) DOI= http://arxiv.org/pdf/0706.1062v2.

[4] Malkiel, B. 2003. *The Efficient Market Hypothesis and Its Critics*. Princeton University. DOI= http://www.princeton.edu/ceps/workingpapers/91malkiel.pdf

[5] Leskovec, J., et. al 2008. *Statistical Properties of Community Structure in Large Social and Information Networks*. Carnegie Mellon University. DOI= http://cs.stanford.edu/people/jure/pubs/ncp-www08.pdf

[6] Leskovec, J., Huttenlocher, D., Kleinberg, J. 2010. *Signed Networks in Social Media*. In Proc. CHI, 2010.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/964696.964697.

[8] Y. G. Flake, K. Tsioutsiouliklis, R.E. Tarjan. *Graph Clustering Techniques based on Minimum Cut Trees*. Technical Report 2002-06, NEC, Princeton, NJ, 2002. DOI= http://www.cs.princeton.edu/~kt/tech02.ps

[9] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. *Trawling the web for emerging cyber-communities*. In Proc. WWW, 1999. DOI= http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html.

[10] M.E.J. Newman, M. Girvan. *Finding and evaluating community structure in networks*. Phys. Rev. E 69, 026113, 2004. DOI= http://arxiv.org/abs/cond-mat/0308217/

[11] Stijn van Dongen. *A cluster algorithm for graphs*. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May 2000.
[http://www.cwi.nl/ftp/CWIreports/INS/INS-R0010.ps.Z].