Social Network Analysis Group Project: Community Detection

Ekapope VIRIYAKOVITHYA, Priya VARADARAJAN, Harshit PALIWAL, Yen Chun LIU, Deborah KEWON

Introduction

The goal of community detection in social network is to identify highly connected clusters of individuals by finding the nodes that can be easily grouped. Community detection is very important because it helps us have an in-depth understanding of false links and spreading processes in various environments, including but not limited to, epidemics and rumor spread. In this paper, we will use walktrap, modularity optimization and edge betweenness algorithms to find possible divisions of a network. As for evaluation metrics, we will use Variation of Information (VI), rand index and adjusted rand index.

Data Source

The data, we are going to use, is the domestic supply of industries in 2012 (US) retrieved from the US Department of Commerce. It contains 405 observations for every individual industry. In order to analyze the communities extensively, the data will be divided into three different grouping schemes:

- 405 Individual Commodities
- 70 Subsectors
- 23 Sectors

Methodological Approach

Our goal is to get the number of communities, information of sectors per community and modularity score in the structure. We will also be analyzing the communities based on every sector and how they are divided based on different algorithms. As mentioned earlier, the communities can be detected using the following algorithms

WalkTrap: It detects communities through a series of short random walks. The idea is that the vertices encountered on any given random walk are more likely to be within a community.

Modularity Optimization (Louvain Modularity) : It detects small communities by optimizing modularity on all nodes. In other words, it joins the pair of communities that most increases modularity until no such pair exists. It is one of the fastest modularity-based algorithms that performs well with large graphs.

Edge-betweenness: It detects communities by iteratively including and removing nodes in order to maximize the modularity.

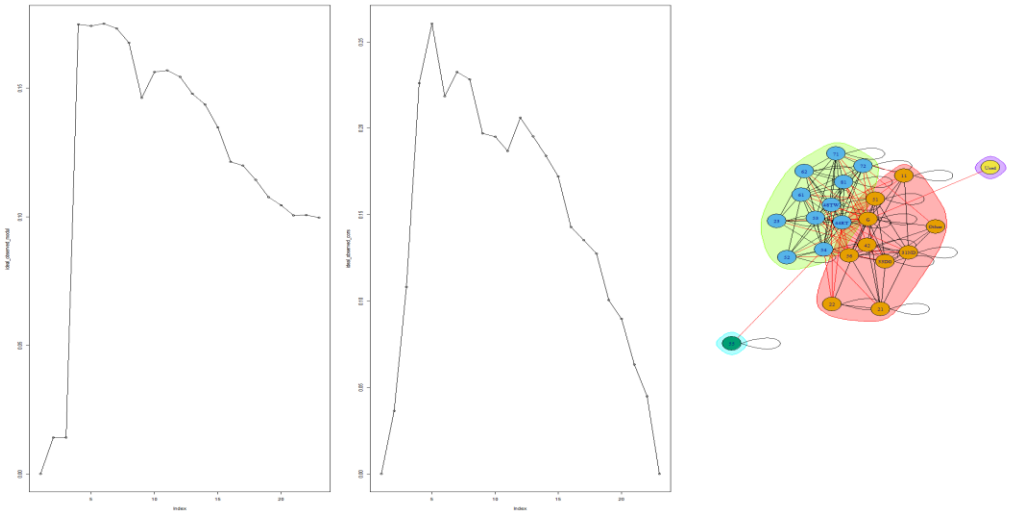
Results

Method	# communities	405 industries	70 sub-sectors	23 sectors
Walktrap	modularity	12	5	4
Modularity	modularity	5	3	5
Edge Betweenness	modularity	20	29	14

Communities and correlation plot for each scheme

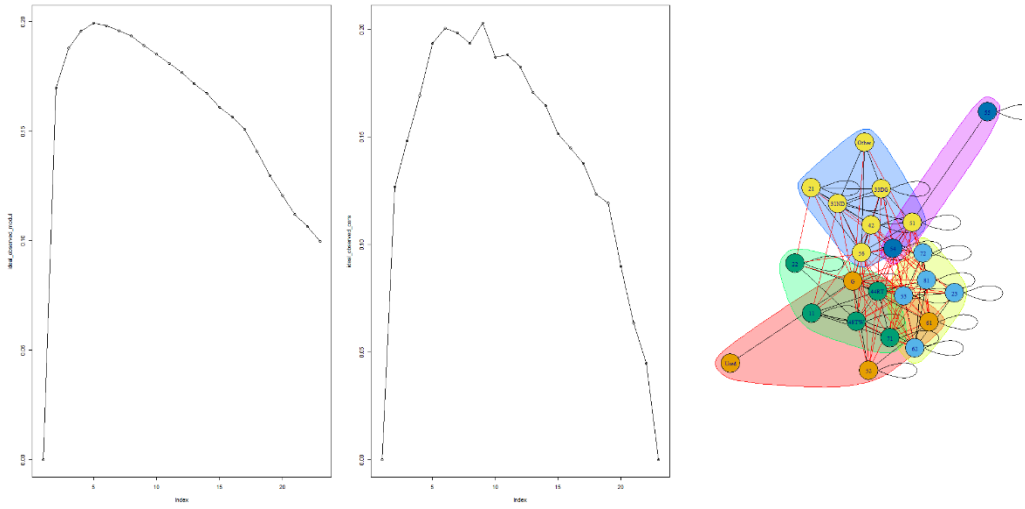
23 Sectors

- **Walktrap**



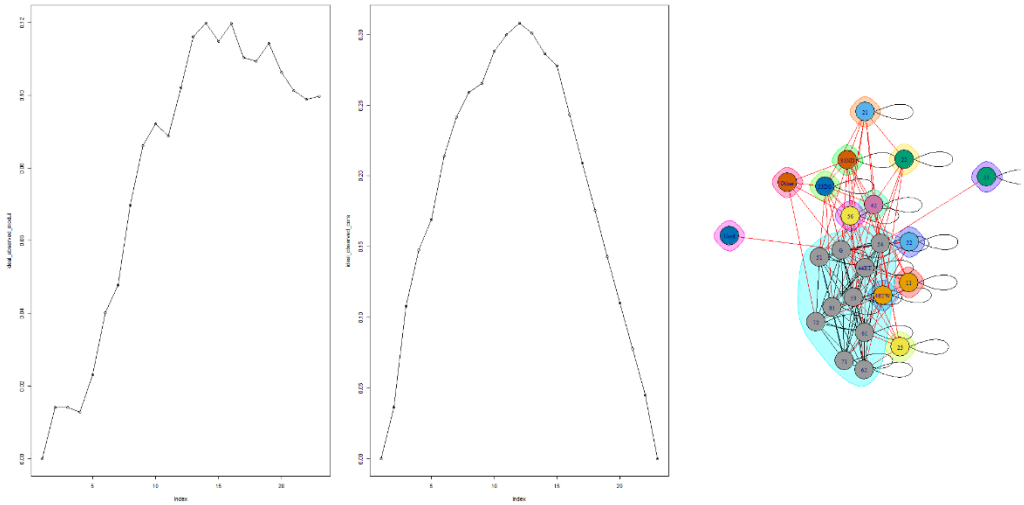
For Walktrap algorithm, the peak in the modularity plot, which is the optimum number of communities is equal to 4. The plot on the right shows us that there are 4 communities detected.

- **Greedy modularity optimization**



In the modularity plot, the peak for modularity optimization algorithm is at 5. The right plot shows us that there are 5 communities detected.

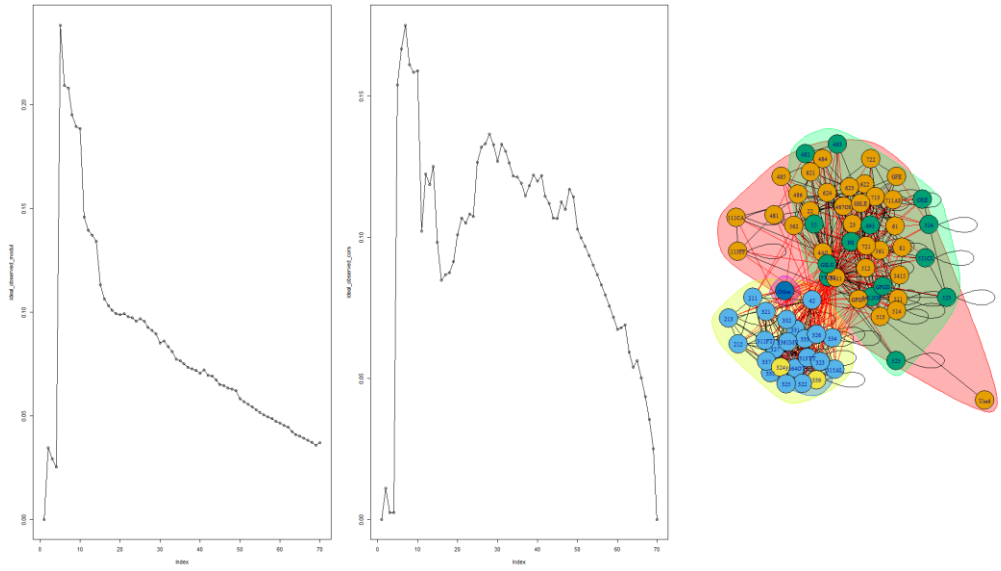
- **Edge-betweenness optimization**



In the above modularity plot for Edge Betweenness algorithm, the peak is at 14. The right graph shows us that there are 14 communities detected.

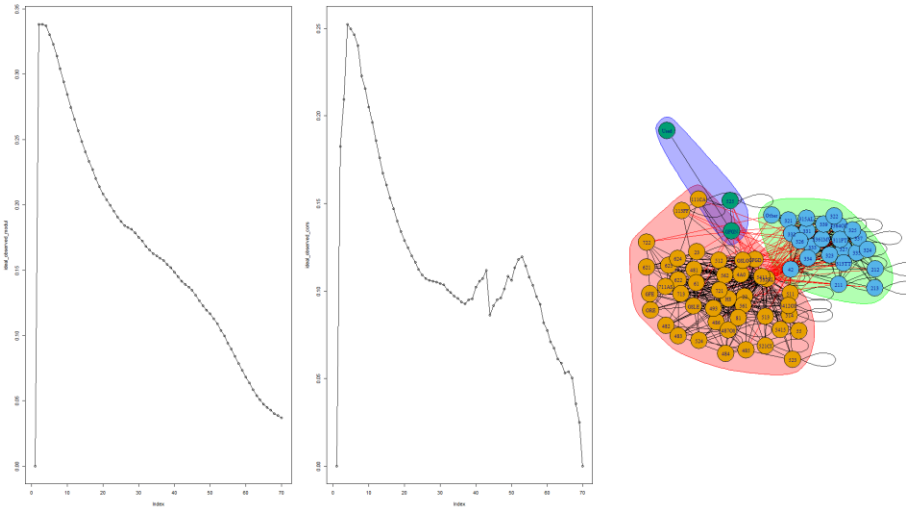
70 subsectors

- **Walktrap**



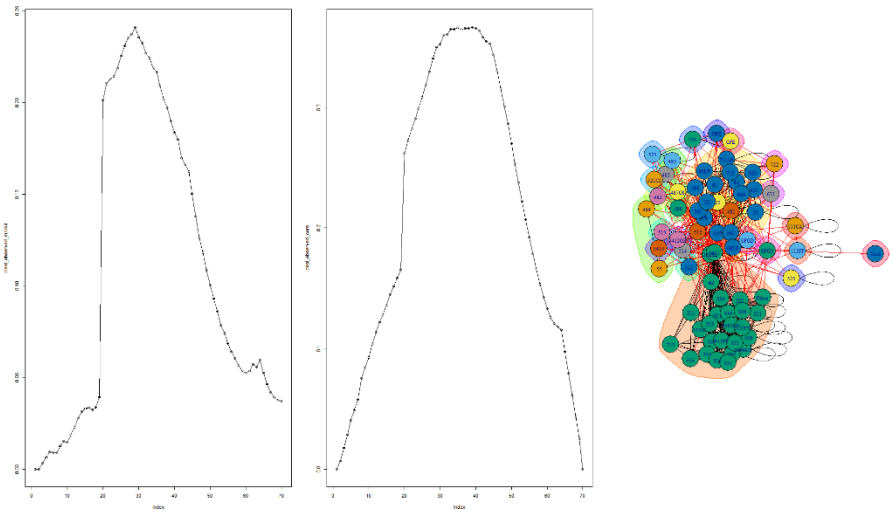
The above plot with modularity for walktrap optimization algorithm shows us that there are 5 communities at the peak for the different sub sectors.

- **Modularity optimization**



The above plot with modularity for Modularity optimization algorithm shows us that there are 3 communities at the peak for the different sub sectors.

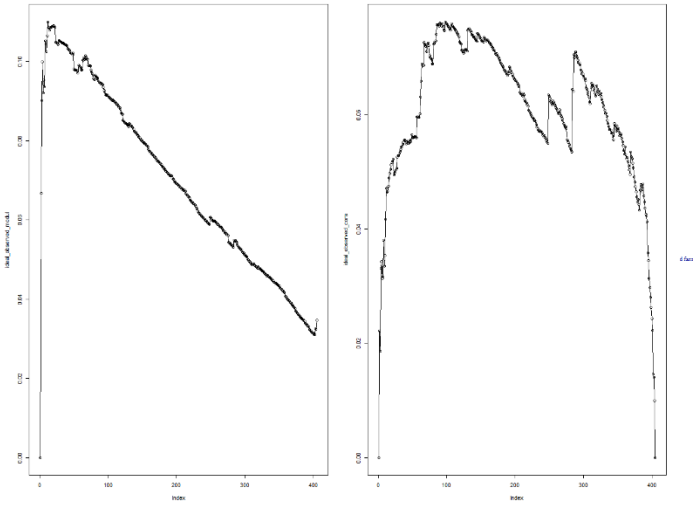
- **Edge-betweenness**



The above plot of modularity for Edge-betweenness algorithm shows us that there are 29 communities at the peak for the different sub sectors.

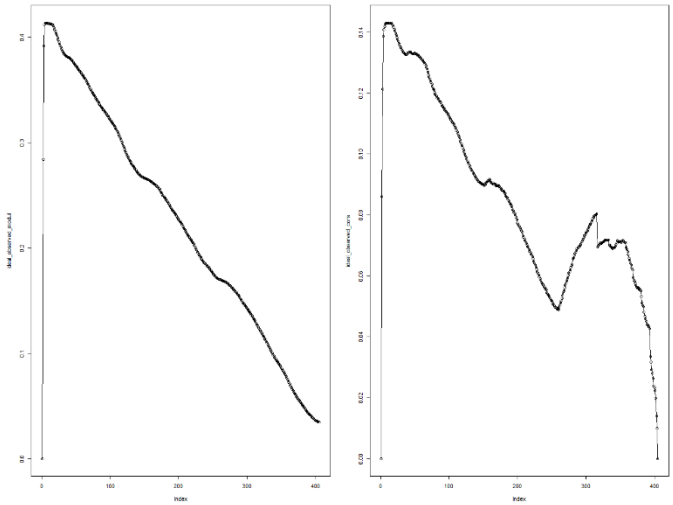
405 industries

- **Walktrap**



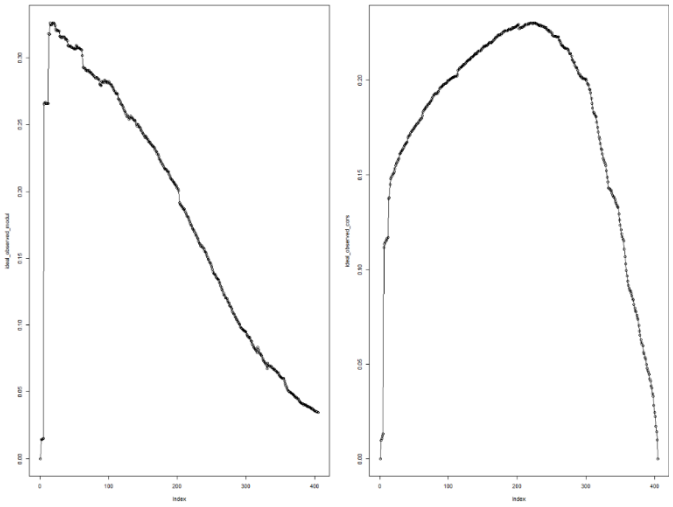
Walk trap has around 12 communities at the peak of modularity with different commodities grouped together.

- **Greedy modularity optimization**



Greedy modularity has around 5 communities at the peak of modularity with different commodities grouped together.

- **Edge-betweenness optimization**



Edge betweenness has around 20 different communities at the peak for the different industries.

Evaluation

Model evaluation is a crucial step in order to determine which model gives better results. There are various evaluation metrics, which are used in order to determine the performance.

For benchmarking the models, a dataset in which we already know the communities are used to check how is the model working against the true values. To evaluate the model, following metrics are used -

- Confusion Matrix
- F-score

After benchmarking the model and finding the best suited model, the model is applied on a similar dataset. This gives optimum results for finding new communities in a similar network.

On the other hand, to evaluate the models which don't have any labeled communities available, following measures are used -

- RAND
- Variation of information

- Adjusted RAND

- **RAND**

  The Rand index/Rand measures a measure of the similarity between two data clustering. A form of the Rand index may be defined that is adjusted for the chance grouping of elements, this is the adjusted Rand index. From a mathematical standpoint, Rand index is related to the accuracy, but is applicable even when class labels are not used. The Rand index has a value between 0 and 1, with 0 indicating that the two data clustering do not agree on any pair of points and 1 indicating that the data clustering is exactly the same.

- **Adjusted RAND**

  The expected value of the Rand Index of two random partitions does not take a constant value (e.g. zero). Thus, Adjusted RAND assumes a generalized hypergeometric distribution as null hypothesis: the two clusters are drawn randomly with a fixed number of clusters and a fixed number of elements in each cluster. Then the adjusted Rand Index is the (normalized) difference of the Rand Index and its expected value under the null hypothesis.

- **Variation of information**

  Variation of information or *shared information distance* is a measure of the distance between the two clusters. It is derived from the *mutual information*, but it is a true metric, *i.e.* it is symmetric and satisfies the triangle inequality.

After applying Walktrap, Modularity Optimization and Edge-betweenness algorithms, the performance of these models was compared to each other using the RAND, Adjusted RAND and VI method. These performance indices show how similar these models are compared to each other.

Below are the results for different model performances -

23 Sectors scheme

- Walktrap and Modularity Optimization

Method	**Performance score**
RAND	0.189
Adjusted RAND	6.850e-17
Variation of Information	1.542

- Modularity Optimization and Edge-Betweenness

Method	Performance score
RAND	0.695
Adjusted RAND	-0.014
Variation of Information	1.871

- Walktrap and Edge Betweenness

Method	Performance score
RAND	0.177
Adjusted RAND	-6.752078e-17
Variation of Information	2.134

70 Sub-sectors scheme

- Walktrap and Modularity Optimization

Method	Performance score
RAND	0.785
Adjusted RAND	0.557
Variation of Information	0.718

- Modularity Optimization and Edge-Betweenness

Method	Performance score
RAND	0.655
Adjusted RAND	0.301
Variation of Information	1.784

- Walktrap and Edge Betweenness

Method	Performance score
RAND	0.726
Adjusted RAND	0.304
Variation of Information	1.995

405 Industries scheme

- Walktrap and Modularity Optimization

Method	Performance score
RAND	0.622
Adjusted RAND	0.056
Variation of Information	2.414

- Modularity Optimization and Edge-Betweenness

Method	Performance score
RAND	0.758
Adjusted RAND	0.517
Variation of Information	0.998

- Walktrap and Edge Betweenness

Method	Performance score
RAND	0.602
Adjusted RAND	0.094
Variation of Information	2.253

Based on the above results, much of inference cannot be made because we do not have the base model for comparison, however we can interpret that Modularity optimization gives us better results on comparison with any algorithm for all the evaluation metrics.

Business Interpretation

In order to further interpret our analysis on commodity detection for the supply chain data from business perspective, we have looked further in the communities that the algorithm provides for all the different types of data (23, 70 and 405 commodities).

The tables below show the detail grouping by each algorithm. For each group, the industries have been provided with similar commodities. Let us take the example of 23 sectors scheme and explain further, we can see 'Finance and insurance' and 'Real estate and Rental and Leasing' have been group together in group 2, this means that these two sectors have provided similar commodities. In the market of rental and leasing, they do have relationships with insurance industries. They tend to collaborate with cross-selling marketing strategy.

For the ones, which are unable to be explained, we may need to go deeper to businesses or seeking for more external information. The use of this information from the marketing side can provide cross-selling within two or multiple industries. By looking at the overall sectors, it could be understood all possible potential markets or collaborated partners, and even M&A (merge and acquisition) plan.

23 sectors scheme

Walktrap

Group	Name
1	AGRICULTURE, FORESTRY, FISHING, AND HUNTING, MINING, UTILITIES, DURABLE GOODS NONDURABLE GOODS, WHOLESALE TRADE, INFORMATION, ADMINISTRATIVE AND WASTE SERVICES, GOVERNMENT, NONCOMPARABLE IMPORTS AND REST-OF-THE-WORLD ADJUSTMENT
2	CONSTRUCTION, RETAIL TRADE, TRANSPORTATION AND WAREHOUSING, EXCLUDING POSTAL SERVICE, FINANCE AND INSURANCE, REAL ESTATE AND RENTAL AND LEASING, PROFESSIONAL AND TECHNICAL SERVICES, EDUCATIONAL SERVICES HEALTH CARE AND SOCIAL ASSISTANCE, ARTS, ENTERTAINMENT, AND RECREATION ACCOMMODATION AND FOOD SERVICES, OTHER SERVICES, EXCEPT GOVERNMENT
3	MANAGEMENT OF COMPANIES AND ENTREPRISES
4	SCRAP, USED AND SECONDHAND GOODS

Greedy Modularity

Group	Sectors
1	FINANCE AND INSURANCE, EDUCATIONAL SERVICES, GOVERNMENT and SCRAP, USED AND SECONDHAND GOODS
2	CONSTRUCTION, REAL ESTATE AND RENTAL AND LEASING, HEALTH CARE AND SOCIAL ASSISTANCE, ACCOMMODATION AND FOOD SERVICES, OTHER SERVICES, EXCEPT GOVERNMENT
3	AGRICULTURE, FORESTRY, FISHING, AND HUNTING, UTILITIES, RETAIL TRADE, TRANSPORTATION AND WAREHOUSING, EXCLUDING POSTAL SERVICE, ARTS, ENTERTAINMENT, AND RECREATION
4	MINING, DURABLE GOODS, NONDURABLE GOODS, WHOLESALE TRADE, INFORMATION, ADMINISTRATIVE AND WASTE SERVICES, NONCOMPARABLE IMPORTS AND REST-OF-THE-WORLD ADJUSTMENT
5	PROFESSIONAL AND TECHNICAL SERVICES, MANAGEMENT OF COMPANIES AND ENTREPRISES

Edge-betweenness optimization

Group	
1	AGRICULTURE, FORESTRY, FISHING, AND HUNTING
2	MINING
3	UTILITIES
4	CONSTRUCTION
5	DURABLE GOODS
6	NONDURABLE GOODS
7	WHOLESALE TRADE
8	RETAIL TRADE, INFORMATION, REAL ESTATE AND RENTAL AND LEASING, PROFESSIONAL AND TECHNICAL SERVICES, EDUCATIONAL SERVICES, HEALTH CARE AND SOCIAL ASSISTANCE, ARTS, ENTERTAINMENT, AND RECREATION, ACCOMMODATION AND FOOD SERVICES, OTHER SERVICES, GOVERNMENT
9	TRANSPORTATION AND WAREHOUSING, EXCLUDING POSTAL SERVICE
10	FINANCE AND INSURANCE
11	MANAGEMENT OF COMPANIES AND ENTREPRISES
12	ADMINISTRATIVE AND WASTE SERVICES
13	SCRAP, USED AND SECONDHAND GOODS
14	NONCOMPARABLE IMPORTS AND REST-OF-THE-WORLD ADJUSTMENT

70 Sectors

Greedy modularity optimization

Group	Sector name
1	Crop production, Forestry, fishing, and related activities UTILITIES, CONSTRUCTION, Other retail, Air transportation, Rail transportation Water transportation, Truck transportation, Transit and ground passenger transportation Pipeline transportation, Other transportation and support activities Warehousing and storage, Publishing industries, except internet (includes software) Motion picture and sound recording industries, Broadcasting and telecommunications Data processing, internet publishing, and other information services Federal Reserve banks, credit intermediation, and related activities Securities, commodity contracts, and investments Insurance carriers and related activities, Housing, Other real estate Rental and leasing services and lessors of intangible assets, Legal services Computer systems design and related services, Miscellaneous professional, scientific, and technical services, Management of companies and enterprises Administrative and support services, Waste management and remediation services Educational services, Ambulatory health care services, Hospitals Nursing and residential care facilities, Social assistance Performing arts, spectator sports, museums, and related activities Amusements, gambling, and recreation industries, Accommodation Food services and drinking places, Other services, except government Federal general government (defense), Federal government enterprises Government, State and local government enterprises
2	Oil and gas extraction, Mining, except oil and gas, Support activities for mining Wood products, Nonmetallic mineral products, Primary metals, Fabricated metal products Machinery, Computer and electronic products, Electrical equipment, appliances, and Components, Motor vehicles, bodies and trailers, and parts, Other transportation equipment Furniture and related products, Miscellaneous manufacturing, Food and beverage and tobacco Products, Textile mills and textile product mills, Apparel and leather and allied products Paper products, Printing and related support activities, Petroleum and coal products Chemical products, Plastics and rubber products, WHOLESALE TRADE, NONCOMPARABLE IMPORTS AND REST-OF-THE-WORLD ADJUSTMENT
3	Funds, trusts, and other financial vehicles, Federal general government (nondefense) SCRAP, USED AND SECONDHAND GOODS

Walktrap

Group	Sector name
1	Crop production, Forestry, fishing, and related activities, UTILITIES, CONSTRUCTION, Other retail, Air transportation, Truck transportation, Transit and ground passenger transportation, Pipeline transportation, Other transportation and support activities, Publishing industries, except internet (includes software), Motion picture and sound recording industries, Broadcasting and telecommunications Data processing, internet publishing, and other information services Legal services, Computer systems design and related services, Administrative and support services, Waste management and remediation services, Educational services Ambulatory health care services, Hospitals, Nursing and residential care facilities Social assistance, Performing arts, spectator sports, museums, and related activities Amusements, gambling, and recreation industries, Accommodation Food services and drinking places, Other services, except government, Federal general government (nondefense), Federal government enterprises, State and local government enterprises, SCRAP, USED AND SECONDHAND GOODS
2	Oil and gas extraction, Mining, except oil and gas, Support activities for mining Wood products, Nonmetallic mineral products, Primary metals Fabricated metal products, Machinery Computer and electronic products, Electrical equipment, appliances, and components Motor vehicles, bodies and trailers, and parts, Other transportation equipment Furniture and related products, Food and beverage and tobacco products Textile mills and textile product mills, Apparel and leather and allied products Paper products, Printing and related support activities, Chemical products Plastics and rubber products, WHOLESALE TRADE
3	Rail transportation, Water transportation, Warehousing and storage Federal Reserve banks, credit intermediation, and related activities Securities, commodity contracts, and investments, Insurance carriers and related activities Funds, trusts, and other financial vehicles, Housing, Other real estate Rental and leasing services and lessors of intangible assets, Miscellaneous professional, scientific, and technical services, Management of companies and enterprises, Federal general government (defense), Government
4	Miscellaneous manufacturing, Petroleum and coal products
5	NONCOMPARABLE IMPORTS AND REST-OF-THE-WORLD ADJUSTMENT

Edge-betweenness optimization

Group	Name of the sector
1	Crop production
2	Forestry, fishing, and related activities
3	Oil and gas extraction, Mining, except oil and gas, Support activities for mining Wood products, Nonmetallic mineral products, Primary metals Fabricated metal products, Machinery, Computer and electronic products Electrical equipment, appliances, and components, Motor vehicles, bodies and trailers, and parts, Other transportation equipment, Furniture and related products Miscellaneous manufacturing, Food and beverage and tobacco products Textile mills and textile product mills, Apparel and leather and allied products Paper products, Printing and related support activities, Petroleum and coal products Chemical products, Plastics and rubber products, WHOLESALE TRADE Rental and leasing services and lessors of intangible assets Legal services, NONCOMPARABLE IMPORTS AND REST-OF-THE-WORLD ADJUSTMENT
4	UTILITIES
5	CONSTRUCTION, Other retail, Warehousing and storage, Housing Administrative and support services, Waste management and remediation services Educational services, Hospitals, Nursing and residential care facilities Social assistance, Performing arts, spectator sports, museums, and related activities Amusements, gambling, and recreation industries, Accommodation Other services, except government, Government, State and local government enterprises
6	Air transportation
7	Rail transportation
8	Water transportation
9	Truck transportation, Management of companies and enterprises
10	Transit and ground passenger transportation
11	Pipeline transportation
12	Other transportation and support activities
13	Publishing industries, except internet (includes software)
14	Motion picture and sound recording industries
15	Broadcasting and telecommunications
16	Data processing, internet publishing, and other information services
17	Federal Reserve banks, credit intermediation, and related activities
18	Securities, commodity contracts, and investments
19	Insurance carriers and related activities
20	Funds, trusts, and other financial vehicles
21	Other real estate
22	Computer systems design and related services
23	Miscellaneous professional, scientific, and technical services
24	Ambulatory health care services
25	Food services and drinking places
26	Federal general government (defense)
27	Federal general government (nondefense)
28	Federal government enterprises
29	SCRAP, USED AND SECONDHAND GOODS

Conclusion

Comparing the results of the three different algorithms, we can say that the modularity optimization algorithm seems very consistent on comparison across the all three different types of data grouping schemes (405 industries, 70 Subsectors and 23 Sectors), while the other two algorithms give different results on different sets of data. However, based on the discussed business interpretation results, it comes to the conclusion that modularity optimization is chosen for the dataset in this project.

References

https://sna.stanford.edu/lab.php?l=3

https://www-complexnetworks.lip6.fr/~latapy/Publis/communities.pdf

http://jgaa.info/accepted/2006/PonsLatapy2006.10.2.pdf

https://www.comp.nus.edu.sg/~leonghw/Courses/CS3230R/Talks/W10-Duy-CS3230R.pdf

https://stackoverflow.com/questions/52717240/cluster-walktrap-returns-three-communities-but-when-plotting-they-are-all-on-to

https://stackoverflow.com/questions/24682166/detecting-network-communities-using-walktrap-using-a-large-number-of-steps

https://www.bea.gov/industry/input-output-accounts-data

https://neo4j.com/docs/graph-algorithms/current/algorithms/louvain/