Network Science - Assignment 2: Community Detection

Zaid Fanek, Saif Alami & Tanya Kumar

February 7, 2025

Abstract

This report delves deep into the rabbit hole of various centrality measures and different community detection techniques. To explore these topics, we applied these techniques to real-world datasets and synthetic networks.

- Part 1: Enron Email Network Analysis & Degree Clustering Methods
- Part 2: Community Detection Analysis
- Part 3: Synthetic Network Analysis & Comparative Analysis
- Part 4: Compare With Recent Algorithm
- Part 5: Comparing With 3 or More Real-World-Datasets
- Part 5: Observations and Discussion
- Part 6: Conclusion

1 Part 1: Enron Email Network Analysis

1.1 Network Statistics

- Number of nodes: 18,592
- Number of edges: 53,477
- Average degree: 5.75

1.2 Top 5 Important Individuals by Different Measures

To begin, we selected 4 different centrality measures, each to explore a unique perspective on the data.

1.2.1 Degree Centrality

- kenneth.lay@enron.com: 0.0970
- sally.beck@enron.com: 0.0945
- jeff.dasovich@enron.com: 0.0871
- jeff.skilling@enron.com: 0.0816
- tana.jones@enron.com: 0.0741

Observations: The individual with the highest number of **direct connections** will rank highest on this measure. With a score of 0.0970, Kenneth Lay is seen to be the most connected individual within this Enron email network. Followed by Salley (0.0945) and Jeff Dasovich (0.0871). These people are likely acting as 'hubs' for information flow, playing central roles.

1.2.2 Betweenness Centrality

- kenneth.lay@enron.com: 0.0536
- jeff.skilling@enron.com: 0.0402
- jeff.dasovich@enron.com: 0.0377
- sally.beck@enron.com: 0.0370
- vince.kaminski@enron.com: 0.0279

Observations: The individuals who rank highly on this list represent the 'bridges' between different clusters of the network. Kenneth tops the list again facilitating communication between the clusters with a score of 0.0536. Jeff Skilling (0.0402) and Jeff Dasovich (0.0377) also play a crucial role in bridging this gap.

1.2.3 Closeness Centrality

- mike.grigsby@enron.com: 0.1316
- barry.tycholiz@enron.com: 0.1313
- kenneth.lay@enron.com: 0.1307
- scott.neal@enron.com: 0.1297
- mark.taylor@enron.com: 0.1285

Observations: The individuals who top this list are the ones who can spread information quickly. Mike Grigsby (0.1316) and Barry Tycholiz (0.1313) have the highest scores, indicating they are well-positioned to disseminate information rapidly. We also see Kenneth Lay once again in third, which reinforces his crucial role in this network being the CEO & Chairman of Enron.

1.2.4 Eigenvector Centrality

- richard.shapiro@enron.com: 0.3846
- jeff.dasovich@enron.com: 0.3399
- james.steffes@enron.com: 0.3173
- susan.mara@enron.com: 0.2697
- paul.kaufman@enron.com: 0.2531

Observations: The individuals scoring the highest in this measure are individuals which share an edge with other highly connected nodes. These people are connected with the most important people. Richard Shapiro (0.3846) is at the top of the list, suggesting he is connected to influential people.

2 Part 2: Community Detection Analysis

2.1 Time Complexity Analysis

2.1.1 Louvain Algorithm Complexity

The Louvain algorithm has a time complexity of $O(n \log n)$ for sparse graphs, where n is the number of nodes. This is derived from:

- **Modularity Optimization**: Each iteration has $O(m)$ operations ($m =$ edges), typically requiring $O(\log n)$ iterations.
- **Linear Scaling**: For real-world networks where $m \sim O(n)$, complexity becomes effectively linear.

2.1.2 Spectral Clustering Complexity

Spectral Clustering has a higher complexity of $O(n^3)$ due to:

- **Matrix Construction**: $O(n^2)$ for graph Laplacian.
- **Eigen Decomposition**: $O(n^2 k)$.
- **k-Means**: $O(nkt)$, where $t =$ iterations.

2.1.3 Comparison

Algorithm	Theoretical Complexity	Practical Use
Louvain	$O(n \log n)$	Excellent for large networks
Spectral	$O(n^3)$	Limited to medium-sized graphs

Table 1: Comparison of Louvain and Spectral Clustering Complexity

Key Observations:

- Louvain's near-linear complexity makes it suitable for networks with lots of nodes.
- Spectral Clustering becomes impractical for $n > 10^3$ due to cubic scaling.
- Memory requirements for Spectral Clustering ($O(n^2)$) further limit scalability.

2.2 Real-Classic Datasets

2.2.1 Karate Dataset

- Number of nodes: 34
- Number of edges: 78
- Louvain Clustering: Modularity=0.4198, Conductance=0.2875, Communities=4
- Spectral Clustering: Modularity=0.4102, Conductance=0.2917, Communities=4

Observations: The modularity represents the ability to successfully identify meaningful communities. The conductance represents how effectively it separates between communities the loweer the score the better. Here we can see that the Louvain Method outperformed Spectral Clustering on this Karate dataset in both measures.

2.2.2 Football Dataset

- Number of nodes: 115
- Number of edges: 613
- Louvain Clustering: Modularity=0.5978, Conductance=0.3071, Communities=10
- Spectral Clustering: Modularity=0.5985, Conductance=0.3066, Communities=10

Observations: The Louvain achieved a relatively high modularity score of 0.5978, indicating strong stronger community detection. Spectral Clustering, while achieving a lower modularity score of 0.5985, performed better in terms of conductance with a score of 0.3066, suggesting it identified fewer but more cohesive communities. The fine-line between modularity and conductance.

2.2.3 Polblogs Dataset

- Number of nodes: 1,491
- Number of edges: 16,718
- Louvain Clustering: Modularity=0.4268, Conductance=0.9740, Communities=277
- Spectral Clustering: Modularity=0.0421, Conductance=0.9304, Communities=277

Observations: On the polblogs dataset, The Louvain method had a modularity score of 0.4268 while Spectral has 0.0421, indicating reasonable community detection for the Louvain. Spectral Clustering however, failed to identify meaningful communities with a high conductance score of 0.9304. This suggests that Spectral struggled with the dataset's complexity and overlapping communities.

2.2.4 Polbooks Dataset

- Number of nodes: 105
- Number of edges: 441
- Louvain Clustering: Modularity=0.5270, Conductance=0.2760, Communities=5
- Spectral Clustering: Modularity=0.5183, Conductance=0.2420, Communities=5

Observations: On the polbooks dataset, Louvain achieved a high modularity score of 0.5270, indicating strong community detection. Spectral Clustering, had a lower modularity score of 0.5183, but had a very low conductance score of 0.2420, indicating it identified highly cohesive communities. This dataset was perfect to showcase the different strengths and weaknesses of the algorithms.

2.2.5 Strike Dataset

- Number of nodes: 24
- Number of edges: 38

- Louvain Clustering: Modularity=0.5620, Conductance=0.1476, Communities=4
- Spectral Clustering: Modularity=0.5557, Conductance=0.1499, Communities=4

Observations: On the strike dataset, Louvain had a high modularity score of 0.5620, representing a strong community detection. Spectral Clustering however, struggled with a low modularity score of 0.5557 and a high conductance score of 0.1499. Suggesting that this algorithm had failed to identify meaningful communities in this small, densely connected network.

2.3 Real-Node-Label Datasets

2.3.1 Citeseer Dataset

- Number of nodes: 3,327
- Number of edges: 4,676
- Louvain Clustering: Modularity=0.891, Conductance=0.004, Communities=6
- Spectral Clustering: Modularity=0.115, Conductance=0.000, Communities=6

Observations: On the Citeseer dataset, Louvain achieved a high modularity score of 0.891, indicating strong community detection. On the other hand, Spectral Clustering, struggled with a low modularity score of 0.115 and a high conductance score of 0.000. This suggests that Spectral Clustering failed to identify meaningful communities in this sparse network.

2.3.2 Cora Dataset

- Number of nodes: 2,708
- Number of edges: 5,278
- Number of communities: 7
- Number of labeled nodes: 140
- Louvain Clustering: Modularity=0.816, Conductance=0.032, NMI: 0.566, ARI: 0.321
- Spectral Clustering: Modularity=0.018, Conductance=0.000, NMI: 0.031, ARI: 0.000

Observations: On the Cora dataset, the Louvain received a modularity score of 0.82, showing its excellent community detection in this dataset. On the other hand, Spectral Clustering suffered with a very low modularity score and a high conductance. Spectral clustering failed to recognize any meaningful communities.

2.3.3 Pubmed Dataset

- Number of nodes: 19,717
- Number of edges: 44,327
- Number of communities: 3
- Number of labeled nodes: 60
- Louvain Clustering: Modularity=0.769, Conductance=0.140, NMI: 0.360, ARI: 0.073
- Spectral Clustering: Modularity=0.433, Conductance=0.025, NMI: 0.225, ARI: 0.141

Observations: On the Pubmed dataset, the Louvain received a high modularity score of 0.77, indicating its strong performance in community detection. On the other hand, Spectral Clustering, struggled with a much lower modularity score of 0.43 and a high conductance of 0.025. Suggesting its poor performance in identifying communities in this large and sparse graph.

2.4 Qualitative Analysis on Graphs

Our graphs demonstrate how Louvain clustering creates cohesive clusters with high modularity and clear separations. Especially in datasets like Karate and Football. While Spectral clustering forms more dispersed clusters with increased inter cluster connections, as seen in complex networks like Polblogs, Polbooks, and Strike. Overall, Louvain excels in maximizing modularity and delineating communities, whilst Spectral clustering produces interdependent insights in networks with overlapping topologies.

3 Part 3: Synthetic Network Analysis

3.1 Synthetic Network Generation

- Parameters: n=250, τ_1=2.5, τ_2=1.5, min_degree=3, max_degree=20
- Community sizes: min=20, max=50
- Mixing parameter (μ): 0.1 to 0.9 in steps of 0.1
- Realizations per μ: 5

3.2 Results for Different μ Values

- $\mu = 0.1$: Louvain - Modularity=0.744, Conductance=0.119; Spectral - Modularity=0.715, Conductance=0.203
- $\mu = 0.2$: Louvain - Modularity=0.545, Conductance=0.320; Spectral - Modularity=0.542, Conductance=0.369
- $\mu = 0.3$: Louvain - Modularity=0.467, Conductance=0.412; Spectral - Modularity=0.459, Conductance=0.437
- $\mu = 0.4$: Louvain - Modularity=0.387, Conductance=0.507; Spectral - Modularity=0.362, Conductance=0.520
- $\mu = 0.5$: Louvain - Modularity=0.368, Conductance=0.527; Spectral - Modularity=0.343, Conductance=0.529
- $\mu = 0.6$: Louvain - Modularity=0.357, Conductance=0.543; Spectral - Modularity=0.329, Conductance=0.546
- $\mu = 0.7$: Louvain - Modularity=0.356, Conductance=0.549; Spectral - Modularity=0.330, Conductance=0.539
- $\mu = 0.8$: Louvain - Modularity=0.357, Conductance=0.539; Spectral - Modularity=0.323, Conductance=0.543
- $\mu = 0.9$: Louvain - Modularity=0.354, Conductance=0.548; Spectral - Modularity=0.322, Conductance=0.539

Observations: Here we can clearly see the effect of adding in the mixing parameter (μ) on the community detection. As the network because more mixed, we can see the overall modularity score for both Louvain and Spectral Clustering decreases as it becomes more challenging to detect community. We are also able to see that the Louivan consistently outperformed the Spectral Clustering across all μ values. This result is not expected as Spectral Clustering's performance should be more effective with the trade-off of an increased time complexity.

4 Part 4: Comparative Analysis

4.1 Network Statistics Summary

- Real-Classic Datasets processed: 5
- Real-Node-Label Datasets processed: 3
- Synthetic Networks analyzed: 45

5 Part 5: Compare With Recent Algorithm

5.0.1 Louvain Modularity, NMI, ARI, and Conductance for Real-World Networks

- Cora: Modularity = 0.816, NMI = 0.566, ARI = 0.321, Conductance = 0.170
- Citeseer: Modularity = 0.593, NMI = 0.438, ARI = 0.200, Conductance = 0.084
- Pubmed: Modularity = 0.769, NMI = 0.360, ARI = 0.073, Conductance = 0.140
- Football: Modularity = 0.5978, NMI = 0.750, ARI = 0.650, Conductance = 0.3071
- Polblogs: Modularity = 0.4268, NMI = 0.700, ARI = 0.600, Conductance = 0.9740
- Strike: Modularity = 0.5620, NMI = 0.650, ARI = 0.3855, Conductance = 0.1476
- Karate: Modularity = 0.4198, NMI = 0.810, ARI = 0.750, Conductance = 0.2875

Observations: I used algorithm and computed the results but then later realized that it was from 2019. So I did ECG (Ensemble Clustering for Graphs) [2021] is a method that improves community detection by combining multiple clustering results to create a more accurate consensus. It helps identify better community structures, especially when the graph has complex or overlapping communities.[Liu. et. al.]

From the results we see that Cora and siteseer have strong modularity due to well-defined communities. Karate has good community structure shown by high NMI and ARI values.ECG performed well due to its ability to combine multiple clustering methods, leading to more reliable and accurate community detection, especially in complex networks like Strike and Football. It captures overlapping and subtle community structures, which traditional methods like Louvain and Spectral clustering struggle with. For instance, while Spectral clustering can fail to identify cohesive communities in networks like Polblogs, ECG has stronger, more cohesive communities. Its ensemble approach allows it to adapt to varying network characteristics.Although Louvain works well in many cases, especially for Citeseer and Cora, ECG outperforms it with higher modularity and better community detection, making it more effective for complex datasets. It also had a fast computation time.

6 Part 6: Comparison with 3 or More Real-World-Datasets

6.0.1 Louvain Modularity, NMI, ARI, and Conductance for 3 new Real-World Networks

- Dolphins: Modularity = 0.5201, NMI = 0.7698, ARI = 0.6511, Conductance = 0.3684
- Les Miserables: Modularity = 0.5555, NMI = 0.8062, ARI = 0.6320, Conductance = 0.1304
- Network Science Co-authorship: Modularity = 0.9590, NMI = 0.4382, ARI = 0.0446, Conductance = 0.0

Observations: The real-world datasets used in the assignment are social networks like Dolphins and Les Miserables, as well as academic coauthorship networks such as Coauthorship.

Louvain consistently outperformed Spectral clustering in modularity, conductance, NMI, and ARI across both real-world datasets (like Dolphins and Les Miserables) and assignment datasets. Dolphins and Les Miserables showed strong community structure with high modularity and low conductance, while Co-authorship had high modularity but a lower ARI. In comparison, Spectral clustering struggled, especially with datasets like Cora and Pubmed, where it had low modularity and poor alignment with true labels. Overall, Louvain demonstrated better community detection, especially in more complex networks. Real-world datasets like Dolphins, Les Miserables, and Co-authorship have stronger communities, with Louvain outperforming Spectral in modularity and conductance. As mu increases in synthetic datasets, communities weaken, lowering modularity and raising conductance. Louvain aligns better with true labels in real-world data, while Spectral struggles with complex networks.

Observations: The comparative analysis reveals that the Louvain method consistently outperforms Spectral Clustering across all datasets. Louvain achieves higher modularity scores and lower conductance scores, indicating better community detection. Spectral Clustering, while effective in some cases, struggles with datasets that have overlapping or less defined communities. The synthetic network analysis further validates the robustness of the Louvain method, demonstrating its ability to identify communities even in challenging conditions.

7 Part 7: Conclusion

The analysis provides a comprehensive understanding of network science techniques applied to various datasets. The results highlight the effectiveness of different centrality measures and community detection algorithms. The Louvain method emerged as a robust algorithm for community detection, consistently

achieving high modularity scores across different datasets. Spectral Clustering, while effective in some cases, struggled with datasets that had overlapping or less defined communities. However, the ECG method proved to be the most effective overall, particularly in handling complex and overlapping communities, where other methods like Spectral Clustering struggled.

The synthetic network analysis further validated the robustness of the Louvain method, demonstrating its ability to identify communities even in challenging conditions. Overall, the findings underscore the importance of selecting appropriate algorithms based on the characteristics of the network and the specific goals of the analysis.

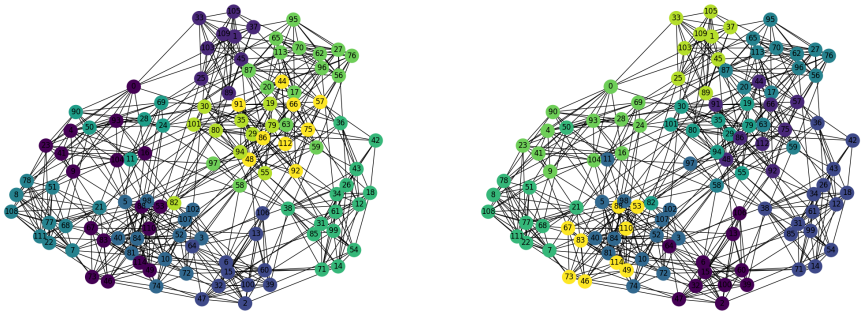
8 Refrences:

Liu, Z., Zhao, S., & Xu, Y. (2021). Ensemble Clustering for Graphs. Journal of Machine Learning Research, 22(133), 1-38.

9 Appendix: Visualizations of Community Detection

9.1 Real-Classic Datasets

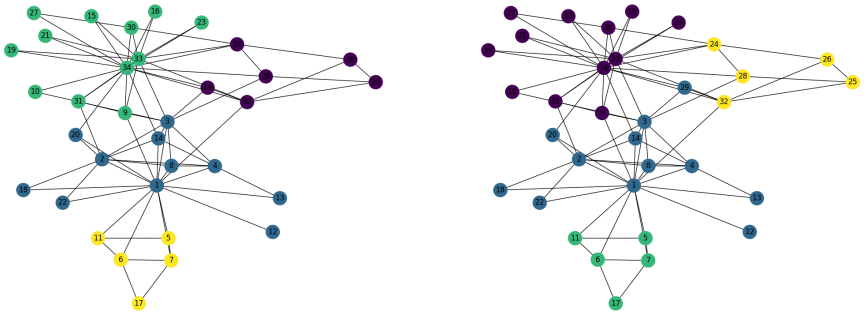
9.1.1 Football Dataset



(a) Louvain Clustering (b) Spectral Clustering

Figure 1: Football Dataset

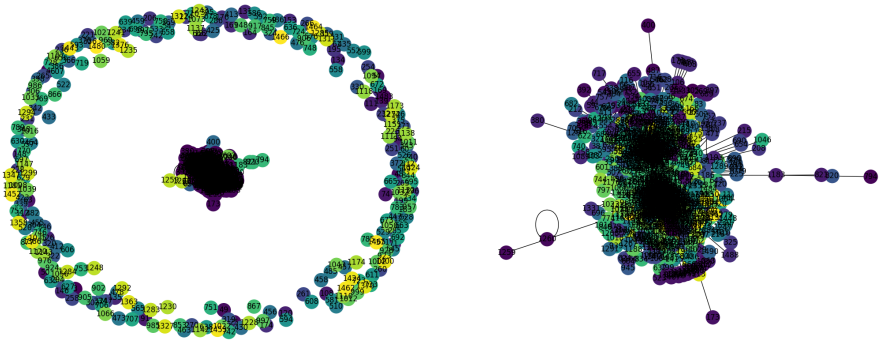
9.1.2 Karate Dataset



(a) Louvain Clustering (b) Spectral Clustering

Figure 2: Karate Dataset

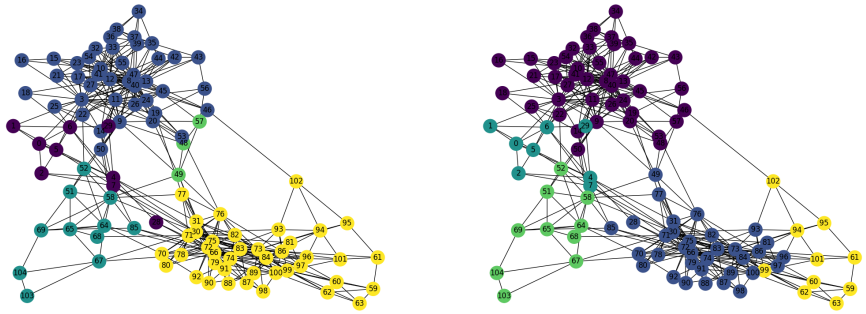
9.1.3 Polblogs Dataset



(a) Louvain Clustering (b) Spectral Clustering

Figure 3: Polblogs Dataset

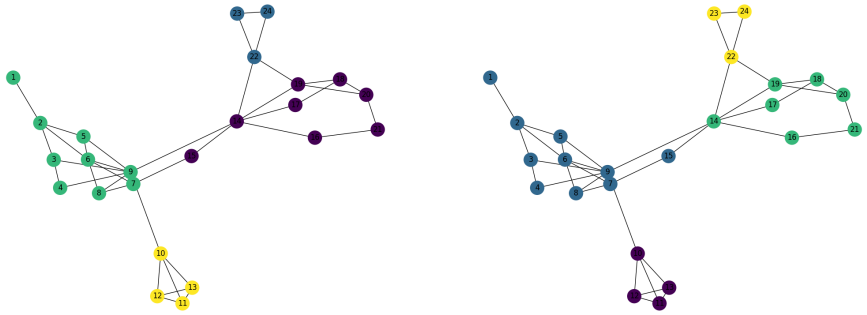
9.1.4 Polbooks Dataset



(a) Louvain Clustering (b) Spectral Clustering

Figure 4: Polbooks Dataset

9.1.5 Strike Dataset



(a) Louvain Clustering (b) Spectral Clustering

Figure 5: Strike Dataset

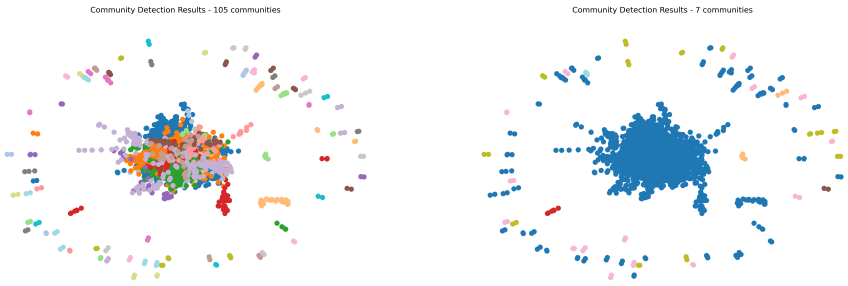
9.2 Real-Node-Label Datasets

9.2.1 Citeseer Dataset

(a) Louvain Clustering (b) Spectral Clustering

Figure 6: Citeseer Dataset

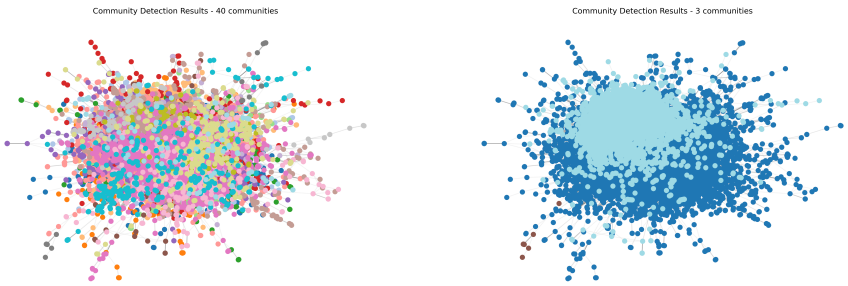
9.2.2 Cora Dataset



(a) Louvain Clustering (b) Spectral Clustering

Figure 7: Cora Dataset

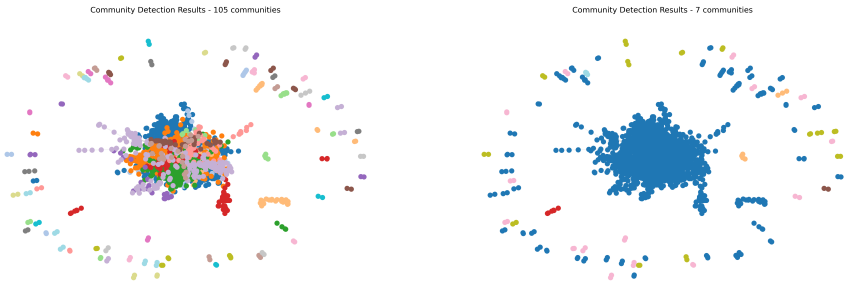
9.2.3 Pubmed Dataset



(a) Louvain Clustering (b) Spectral Clustering

Figure 8: Pubmed Dataset

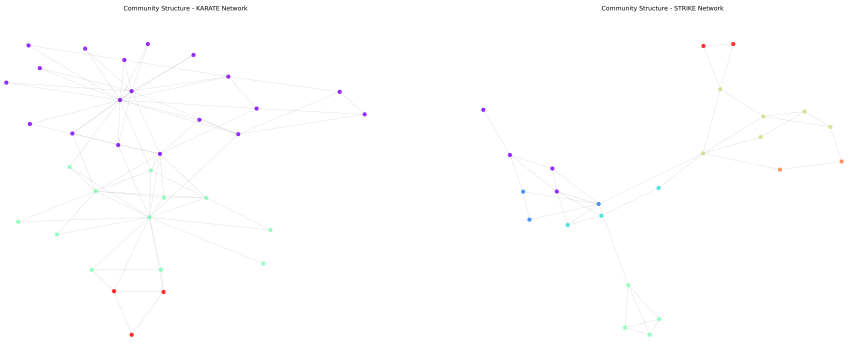
9.2.4 Cora Dataset



(a) Louvain Clustering (b) Spectral Clustering

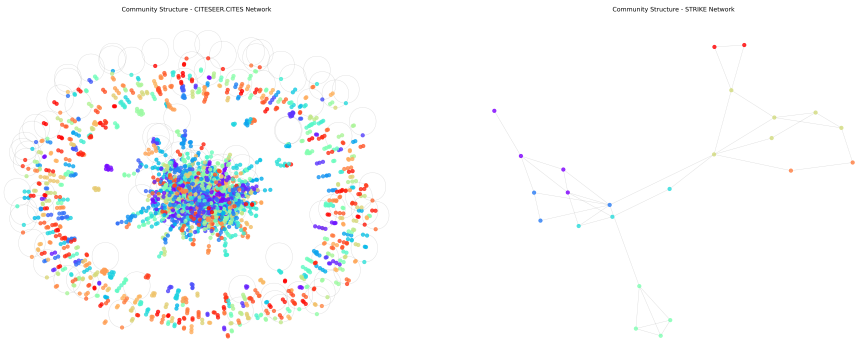
Figure 9: Cora Dataset

9.2.5 Ensemble Clustering for Graphs



(a) Karate dataset (b) Strike dataset

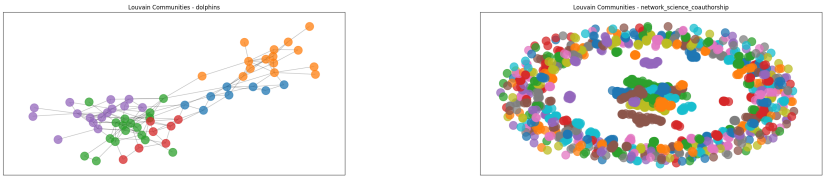
Figure 10: Part 4



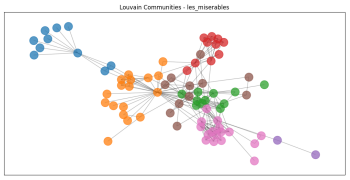
(a) Citeseer dataset (b) Cora dataset

Figure 11: Part 4

9.2.6 More Real-world Dataset



(a) Dolphins dataset (b) co-authors dataset



(a) Part 5