A Study of Community Detection by applying Girvan-Newman Algorithm

Mahurjya Dey  
Department of Computer Science and Enginerring   
The Assam Royal Global UniversityGuwahati,Assam,India  
madhurjyadey45@gmail.com

Chinmoy Hazarika  
Department of Computer Science and Enginerring  
The Assam Royal Global UniversityGuwahati,Assam,India  
chinmoyh12345@gmail.com

Nichol Das  
Department of Computer Science and Enginerring  
The Assam Royal Global UniversityGuwahati,Assam,India  
nicholdas07@gmail.com

*Abstract*—Community detection, a fundamental task for network analysis, aims to partition a network into multiple sub-structures to help reveal their hidden functions. Community detection has been extensively studied and broadly applied to many real-world network problems. Classical approaches to community detection typically utilizes probabilistic graphical models and adopt a variety of prior knowledge to infer community structures. As the problems that network methods try to solve and the network data analyzed become increasingly more sophisticated, new approaches have been proposed and developed, particularly those that utilize deep learning and convert network data into two-dimensional representation. Despite all the recent advancement, there is still a lack of insightful understanding of the theoretical and methodological underpinning of community detection, which will be critically important for future development in the area of network analysis. In this paper, we have taken the datasets from real world networks and have applied Newman-Girvan Algorithm to the identified communities along with the values of Precision, Recall, Accuracy and F1-Score. To visualize each dataset, we have used Gephi to identify the communities and to clearly distinct each community.

Keywords— Newman-Girvan Algorithm, node, graph, communities

# Introduction

Community detection refers to clustering or grouping of nodes within a network in such a way that reveals important patterns or relationships within the network. Many algorithms for community detection have been developed, using techniques and tools from different disciplines such as biology, physics, social sciences, applied mathematics and computer sciences. However, a single community detection algorithm fails to perform in all kind of networks because there is a wide variety of complex networks which are generated from different processes. Community detection algorithms strongly depend on the topology of the networks since they may be static or dynamic. In a static network, community discovery is an easy task as compared to a dynamic network. [6] There exists a number of community detection approaches in static networks, which are mostly optimization-based algorithms that seek an optimal solution according to the defined objective function. Community detection in dynamic networks [7] is a challenging task since such networks are multi-graphs and a pair of nodes can have links appearing or disappearing at different time points. It also leads to constant changes in graph partitions between slices of a multi-graph which make community detection difficult in dynamic networks.

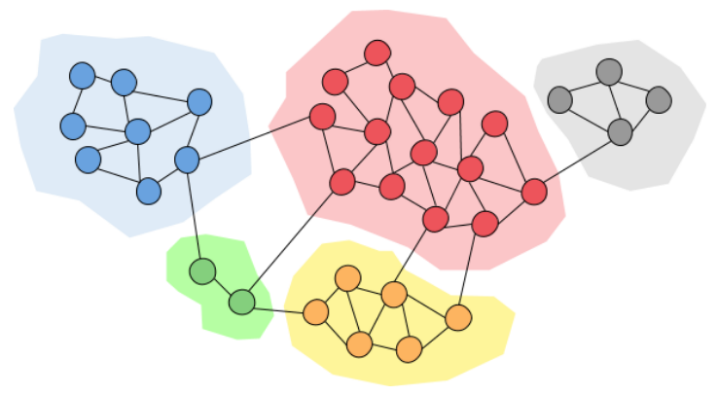


Figure 1: A simple graph with five communities having thirty-four nodes

Above is a diagram to better understand how community detection works, as we can observe from the diagram that it is a graph where different groups are highlighted and each node falls under that desired groups or can be called as community. Such communities are identified by applying different types of algorithms.

As we have applied Newman-Girvan Algorithm upon all the datasets The Girvan–Newman Algorithm [8] detects the communities by progressively removing edges from the original graph. The algorithm removes the “most valuable” edge, traditionally the edge with the highest betweenness centrality, at each step. As the graph breaks down into pieces, the tightly knit community structure is exposed and the result can be depicted as a dendrogram. The same thing can be observed from the above diagram.

We have also used the NetworkX library for study and analysis of complex networks. It provides a diverse collection of real-world networks and is crucial for research in social network analysis, community detection, and graph theory. The datasets vary from one another, offering rich representation of different interaction modalities, making them valuable for testing algorithms and demonstrating network properties.

To evaluate each dataset, we have taken the values of Accuracy that measures the overall correctness of the model by comparing the number of correct predictions to total predictions, Precision that quantifies how many of the predicted connections or community are actually correct, Recall measures the ability to identify all relevant connections or members and F1 score that provides a harmonic mean of precision and recall. [9]

Overall, real-world datasets have been selected to apply the Girvan-Newman algorithm for community detection, aiming to uncover meaningful clusters within complex social networks. This algorithm works by iteratively removing edges with the highest betweenness centrality, effectively breaking the network down into smaller, densely connected communities. Alongside the detection process, performance evaluation metrics such as Accuracy, Precision, Recall, and F1-Score have been calculated as these metrics provide a quantitative means to assess how well the algorithm identifies relevant communities compared to known or expected groupings in the datasets. By comparing these values across different networks, the effectiveness and reliability of the Girvan-Newman algorithm in various real-world contexts can be thoroughly examined, offering insights into its application.

# RELATED WORKS

The Girvan-Newman algorithm is mainly used in community detection; with the help of this algorithm, it helps to reveal group structures in complex networks using an edge betweenness-based strategy. By progressively removing edges with high centrality, the algorithm effectively splits a network into cohesive subgroups, uncovering modular organizations that standard clustering methods may miss.[8] It has the ability to detect meaningful patterns within social, biological, and technological systems, making it a benchmark tool for network analysis.​

Evaluations of community detection methods, including Girvan-Newman, frequently employ metrics such as precision, recall, accuracy, and F1-score to quantitatively assess performance against known or ground truth partitions as these measures help compare algorithmic output to known or expected community structures and guide further methodological development [10]. Further research has focused on addressing the algorithm’s computational intensity on large networks by proposing heuristic or centrality-based enhancements while preserving or improving detection quality.[11],[12] The Girvan-Newman framework’s insights have inspired a broad range of advances and remain relevant to uncover the communities and make communities distinct from one another and also helps into investigating complex, multilayer, and dynamic networks.

Community detection plays a crucial role in understanding social networks and improving recommendation systems. The Girvan-Newman algorithm, a hierarchical divisive approach based on edge betweenness centrality, is widely adopted for identifying meaningful communities within large datasets. Its integration with methods like Alternating Least Squares (ALS) for matrix factorization has been shown effective in addressing challenges like the cold start problem in recommendation systems by extracting communities and leveraging user-item ratings within those communities.[14] Numerous studies have demonstrated the scalability and accuracy improvements when combining community detection algorithms with advanced recommendation models, enabling personalized and robust user recommendations in dynamic social network environments.

The Girvan-Newman algorithm works on different types of randomly generated graphs such as undirected, cyclic directed, and acyclic directed graphs. It assesses execution time and clustering quality for graphs of various sizes. [15]

Community detection has emerged as a primary method for understanding how network structure affects behavioural patterns in many systems [16]. In this respect, communities are recognized as an effective way of identifying the underlying structures inferred from network topology and any additional information about the content of these nodes. This includes, for instance, finding potential friends from social media graph, identifying products that can be recommended to users from some user-product purchase network, and identifying social opinions from user-discussion network [17],[18]. The detection of a community group in the graph captures the tendency of nodes to create the groups, according to the employed similarity metric. Another terminology used for community detection is cluster. Although, clusters and communities are used interchangeably in graph theory literature, clustering sometimes stand for a single modality approach of community detection [19]. The ability to detect communities of a given network contributes to the understanding of the inherent properties and structure of this network. Additionally, it provides an overview of communication among nodes of the same community as well as an analysis of the utility of these nodes [20],[21]. Typically, a community is defined as a collection of nodes that have a high degree of connection among themselves, but a low degree of connection with nodes outside the community set. Especially, community detection involves the use of some optimization strategy for transforming a large-scale complex network into a set of disjoint and compact subgroups, without prior knowledge about the number of subgroups and their sizes.

Several algorithms for community detection have been

developed, combining techniques and tools from different

disciplines such as biology [22], physics [23], statistics [24],

social sciences [25], mathematics [26], economics [27], and computer sciences [28]. It is commonly acknowledged that there is no unique community detection algorithm that can universally accommodate all kinds of social networks with high accuracy because of the discrepancy in network types and purposes. For example, in one type of network, algorithmic biases can increase the efficiency, while in another type, they may decrease the efficiency [29], [30], [31].

# METHODOLOGY

## Newman-Girvan Algorithm

Girvan-Newman method is divisive method where edge weight is the number of shortest paths passing through the edge. That value is called edge betweenness and it is a generalization of central vertex betweenness which determines vertex influence on other vertices in network. Vertex betweenness is the number of shortest paths passing through the vertex, therefore, edge betweenness is the number of shortest paths passing through the endpoints of the edge.

The idea is to find which edges in a network occur most frequently between other pairs of nodes by finding edge betweenness centralities. The edges joining communities are then expected to have a high edge betweenness. The underlying community structure of the network will be much more fine-grained once the edges with the highest betweenness are eliminated which means that communities will be much easier to spot.[13] The Newman-Girvan Algorithm can be divided into 3 steps:

* Calculate the betweenness score for all the edges in the network.
* The edge having the highest edge betweenness score will be removed.
* After removal of the edge, betweenness score will be recalculated for all the remaining edges in the network.
* Step 2 will be repeated until we remove all edges or we get the single node in the network.

## Datasets

Our study involved testing the Newman-Girvan algorithm across three different real-world network structures: Zachary Karate Club, American College Football Network, Bottlenose Dolphin Network, Florentine Families and Napoleon Russian Campaign. A brief summary of each of the five data sources are provided below:

1) Zachary Karate Club: [1] It represents interactions among 34 members of a university karate club studied by Wayne W. Zachary. It captures social ties between pairs of members based on their interactions outside the club activities. A key event that makes this dataset famous is a split in the club due to a conflict between the instructor and the club president, resulting in two communities.

2) American College Football Network: The American College Football Network dataset [2] represents the games played between Division IA college football teams during the regular season. The dataset consists the list of teams present in the Division and also contains the games played between teams within for a season. It contains 115 teams and has recorded 613 games during a season.

3) Bottlenose Dolphin Network: It represents the social associations among [3] bottlenose dolphins in Doubtful Sound, New Zealand. It notes down the patterns of long-lasting social bonds and group structures observed over time. It consists of approximately 62 bottlenose dolphins leaving in Doubtful Sound, New Zealand. It also notes down the social interactions between the dolphins

4) Florentine Families: It represents the social, business, and marriage ties among prominent families of Renaissance Florence [4] during the 14th and 15th centuries The data was compiled by John Padgett. With around 16 families and connections representing marriage and financial ties.

5) Napoleon Russian Campaign: The dataset particularly focuses on 1812 invasion in Russia [5], to understand the complex political and military alliances, conflict dynamics, and coalition formations among European states at the time. The dataset lists out the nations as nodes with diplomatic, military, or alliance ties.

## Experimental Set up

All the programs are coded in python. The execution and the testing are done on a machine with AMD Ryzen 7 6800H and 16 GB of memory.

# RESULTS AND DISSCUSSION

## Zachary Karate Club

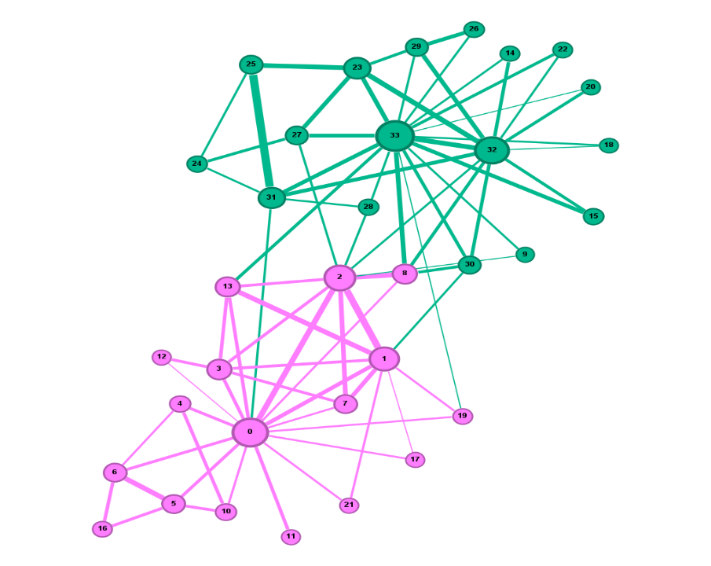


Figure 2: Zachary Karate Club is divided into two communities.

Discussion: It is observe in most of the papers that Zachary Karate Club is divided into two communities where one is Administrators of the club and another is Instructors of the club. A disagreement was raised in both the communities and the Instructors left out from the club and made one new club.

## American College Football Network:



Figure 3: American College Football Network is divided into twelve sub-communities.

Discussion: In this network, we can observe that the node represents team and the edges represents the games played between the teams. There is a total of twelve teams in the network in the above figure.

## Bottlenose Dolphins Network:

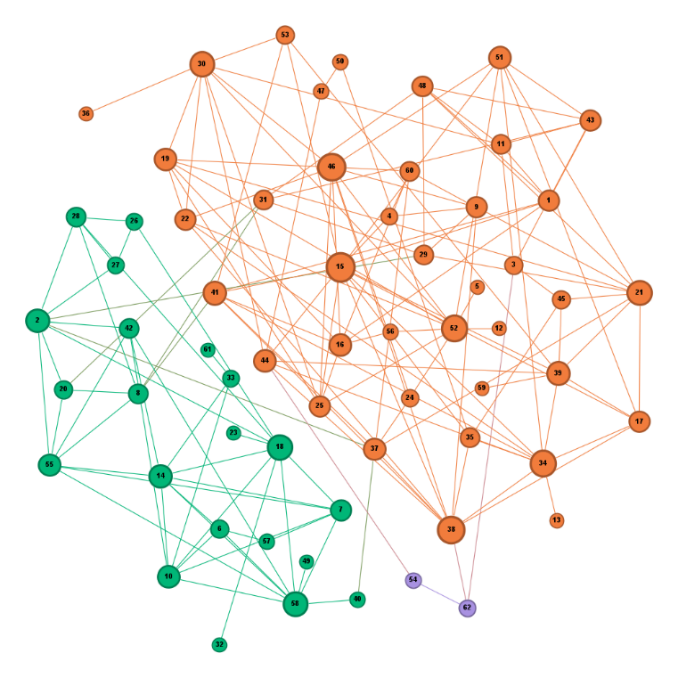


Figure 4: Bottlenose Dolphins Network is divided into two communities.

Discussion: It consists of 62 bottlenose dolphins and all the dolphins have the relation with some another dolphins which is shown in the above figure. There are two major communities in it.

## Florentine Families:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sl No | Datasets | Precision | Recall | Accuracy | F1-Score |
| 1 | Zachary Karate Club | 0.894 | 1 | 0.941 | 0.944 |
| 2 | American College Football Network | 0.894 | 1 | 0.941 | 0.944 |
| 3 | Bottlenose Dolphin Network | 0.717 | 0.67 | 0.90 | 0.661 |
| 4 | Florentine Families | 0.756 | 0.733 | 0.733 | 0.73 |
| 5 | Napoleon Russian Campaign | 0.75 | 1 | 0.84 | 0.857 |

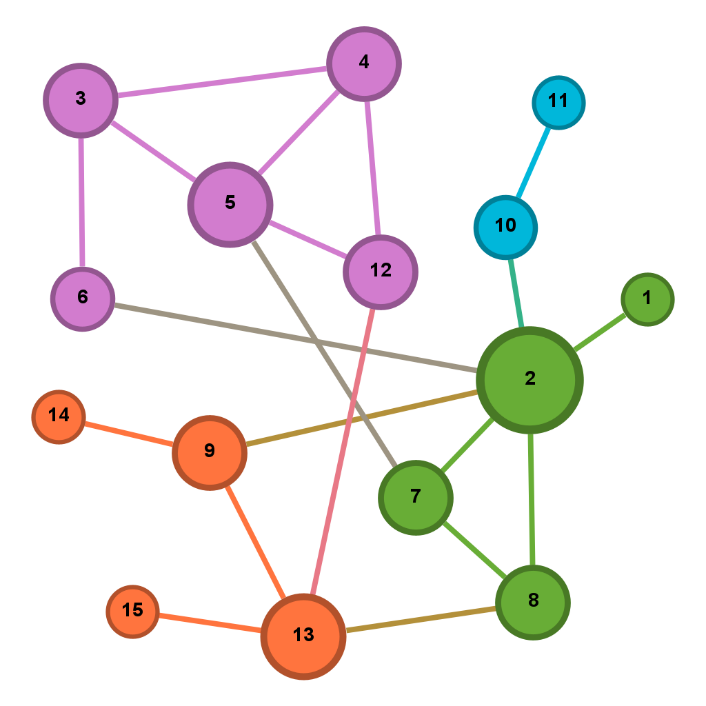


Figure 5: Florentine Families is divided into four sub-communities.

Discussion: In this network, we can observe that different families are nodes) and their business or marriage alliances are the edges. The node size reflects the relative importance or centrality of each family within the network.

## Napoleon Russian Campaign:

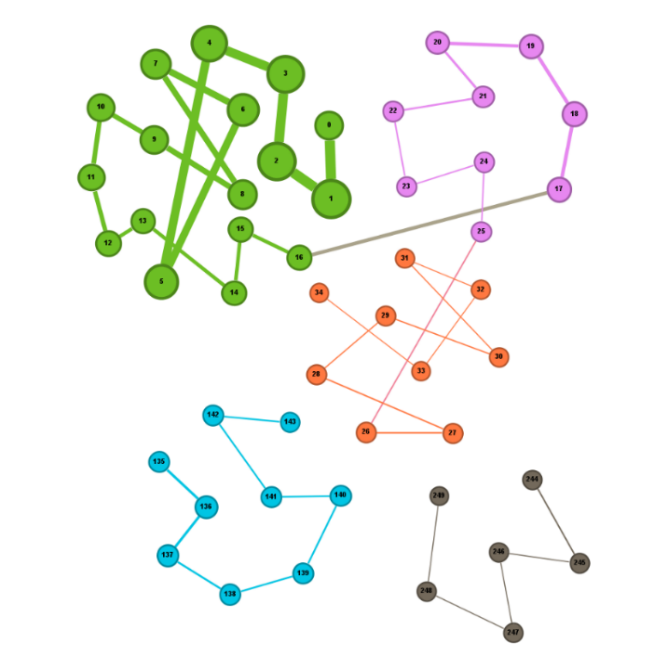


Figure 6: Napoleon Russian Campaign network is divided into five communities.

Discussion: It illustrates the network of alliances and hostilities between nations during Napoleon’s Russian campaign, with each node representing a country or military faction. The communities highlight different coalitions or groups, while the edges depict diplomatic, military, or adversarial relationships. The figure visualizes the fragmentation and connectivity among European states during this historic conflict.

## Result of Precision,Recall,Accuracy and F1-Score of the given Datasets:

Table 1: Values of all the Datasets.

In the above table we have taken the results of the five dataset that has been mention and we can observe that Zachary and American has the highest precision value of 0.894 while the Dolphin has the lowest precision value of 0.717. In the recall section we can observe that Zachary, American and Napolean have the highest value of 1while the dolphin has the lowest recall value of 0.67. In the accuracy section Zachary and American have the highest value of 0.941 and the lowest being the Florentine with the value of 0.73. Lastly, in the F1-Score the highest value is the Zachary and American with a value of 0.944 and the lowest being the Dolphin with a value of 0.661.

## ROC Curve of the given Datasets:

### Zachary Karate Club:

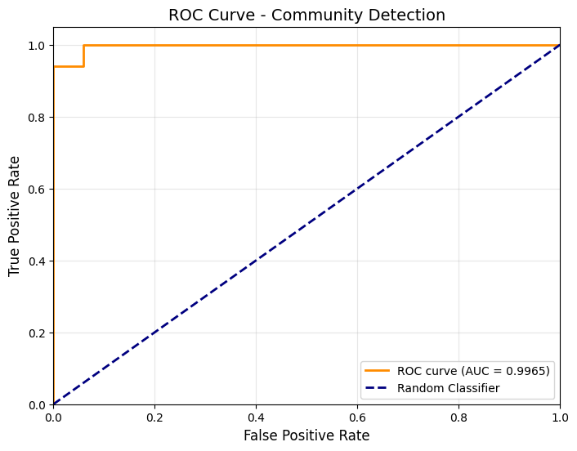


Figure 7: ROC curve for Zachary Karate Club:

Discussion: It shows an excellent classifier performance, with AUC of 0.9965. The curve rises steeply towards the top-left corner, indicating a high true positive rate and a very low false positive rate.

### American College Football Network

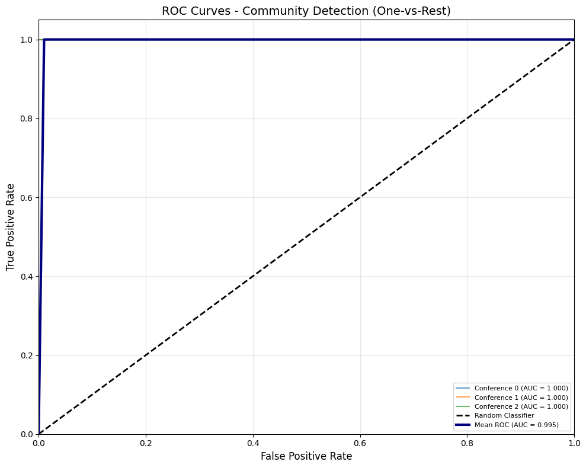


Figure 8: ROC Curve for American College Football Network

Discussion: It has an outstanding classification performance, with each conference showing an AUC of 1.000 and the mean ROC curve achieving an AUC of 0.995. The curves are nearly perfect, hugging the top-left corner, which indicates extremely high true positive rates and very low false positive rates across all classes.

### Bottlenose Dolphins Network:

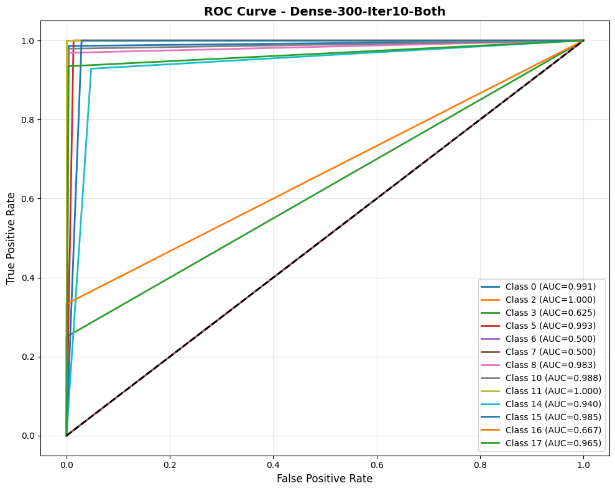


Figure 8: ROC Curve for Bottlenose Dolphins Network

Discussion: It visualizes the community detection performance of a classifier across multiple classes, with each colored line representing a different class. Most curves are very close to the top-left corner, indicating a high true positive rate and low false positive rate for those classes, and AUC scores mostly above 0.98. However, a few classes show lower AUC values, suggesting somewhat weaker distinction for those specific labels.

### Florentine Families:

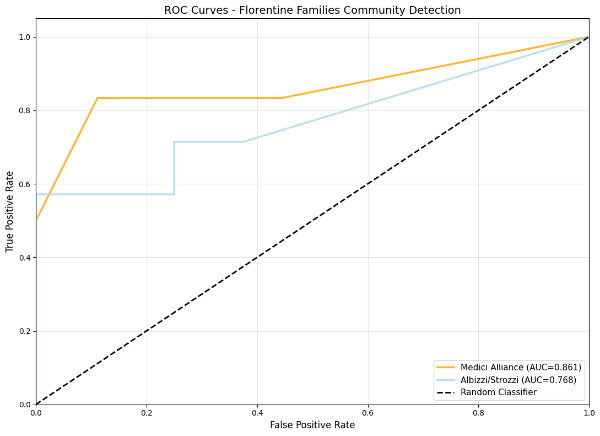


Figure 4: ROC Curve for Florentine Families

Discussion: The AUC score of Medici alliance (AUC = 0.861) and Albizzi/Strozzi group (AUC = 0.766) that shows that Medici alliance curve is closer to the top-left, indicating better performance, while the Albizzi/Strozzi curve lags, showing more challenges in accurately classifying members of this group.

### Napoleon Russian Campaign:

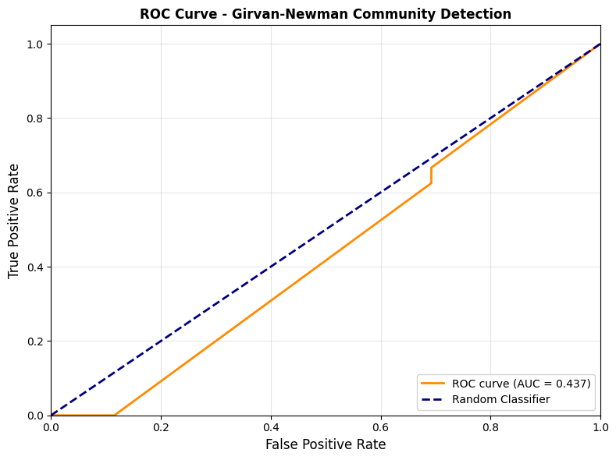


Figure 4: ROC Curve for Napoleon Russian Campaign

Discussion: The area under the curve (AUC) is 0.437, which is below the diagonal line representing a random classifier. It shows weak or ineffective classification quality on this dataset.

# CONCLUSION AND FUTURE SCOPE

We have applied Girvan-Newman to identify community structure in complex networks, offering clear picture of the communities that it has detected from the group of nodes. Its allows the detection of meaningful, hierarchical partitions that align with real-world social and biological groupings. Girvan-Newmann algorithm works well on small to medium-sized networks, but its computational demands pose challenges for large-scale datasets, to solve this problem we need an algorithm which is suitable for larger datasets.

Future work can focus on enhancing algorithm scalability, enabling efficient analysis of massive and dynamic networks that compromise thousands to millions of nodes. By integrating with machine learning and deep learning models automating and refining community detection,

##### References

1. Zachary WW. An information flow model for conflict and fission in small groups. Journal of anthropological research. 1977 Dec 1;33(4):452-73.
2. Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical review E. 2004 Feb 26;69(2):026113.
3. Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait?. Behavioral ecology and sociobiology. 2003 Sep;54(4):396-405.
4. Padgett JF, Ansell CK. Robust Action and the Rise of the Medici, 1400-1434. American journal of sociology. 1993 May 1;98(6):1259-319.
5. Freeman L. The development of social network analysis. A Study in the Sociology of Science. 2004 Jan;1(687):159-67.
6. Javed MA, Younis MS, Latif S, Qadir J, Baig A. Community detection in networks: A multidisciplinary review. Journal of Network and Computer Applications. 2018 Apr 15;108:87-111.
7. Huang, L. C., Yen, T. J., & Chou, S. C. T. (2011, July). Community detection in dynamic social networks: A random walk approach. In 2011 International Conference on Advances in Social Networks Analysis and Mining (pp. 110-117). IEEE.
8. Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.
9. Nasrullah S, Jalali A. [Retracted] Detection of Types of Mental Illness through the Social Network Using Ensembled Deep Learning Model. Computational Intelligence and Neuroscience. 2022;2022(1):9404242.
10. Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. New journal of physics. 2009 Mar 10;11(3):033015.
11. Newman ME. Fast algorithm for detecting community structure in networks. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics. 2004 Jun;69(6):066133.
12. Sathiyakumari K, Vijaya MS. Community detection based on girvan newman algorithm and link analysis of social media. InAnnual Convention of the Computer Society of India 2016 Nov 23 (pp. 223-234). Singapore: Springer Nature Singapore.
13. Choudhury D, Bhattacharjee S, Das A. An empirical study of community and sub-community detection in social networks applying Newman-Girvan algorithm. In2013 1st international conference on emerging trends and applications in computer science 2013 Sep 13 (pp. 74-77). IEEE.
14. Permani AR, Narayan S. Community Detection using Girvan Newman algorithm in Recommendation systems.
15. Matijevic T, Vujicic T, Ljucovic J, Radunovic P, Balota A. Performance Analysis of Girvan-Newman Algorithm on Different Types of Random Graphs. InCentral European Conference on Information and Intelligent Systems 2016 (p. 11). Faculty of Organization and Informatics Varazdin.
16. Zardi H, Alharbi B, Karamti W, Karamti H, Alabdulkreem E. Detection of community structures in dynamic social networks based on message distribution and structural/attribute similarities. IEEE Access. 2021 Apr 29;9:67028-41.
17. He D, Jin D, Baquero C, Liu D. Link community detection using generative model and nonnegative matrix factorization. PloS One. 2014 Jan 28;9(1):e86899.
18. Chen M, Kuzmin K, Szymanski BK. Community detection via maximization of modularity and its variants. IEEE Transactions on Computational Social Systems. 2014 Apr 9;1(1):46-65.
19. Mittal S, Sengupta D, Chakraborty T. Hide and seek: outwitting community detection algorithms. IEEE Transactions on Computational Social Systems. 2021 Mar 12;8(4):799-808.
20. Reichardt J, Bornholdt S. Detecting fuzzy community structures in complex networks with a Potts model. Physical review letters. 2004 Nov 19;93(21):218701.
21. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics. 2006 Jul;74(1):016110.
22. Azadifar S, Rostami M, Berahmand K, Moradi P, Oussalah M. Graph-based relevancy-redundancy gene selection method for cancer diagnosis. Computers in Biology and Medicine. 2022 Aug 1;147:105766.
23. Hric D, Darst RK, Fortunato S. Community detection in networks: Structural communities versus ground truth. Physical Review E. 2014 Dec;90(6):062805.
24. Banks J, Mohanty S, Raghavendra P. Local statistics, semidefinite programming, and community detection. InProceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA) 2021 (pp. 1298-1316). Society for Industrial and Applied Mathematics.
25. D. Liu, G. Yang, Y. Wang, H. Jin, and E. Chen, ‘‘How to protect ourselves from overlapping community detection in social networks,’’ IEEE Trans. Big Data, vol. 8, no. 4, pp. 894–904, Aug. 2022.
26. Zhong Z, Wang X, Qu C, Wang G. Efficient algorithm based on non-backtracking matrix for community detection in signed networks. IEEE Transactions on Network Science and Engineering. 2022 Mar 9;9(4):2200-11.
27. Mansano RE, Allem LE, Del-Vecchio RR, Hoppen C. Balanced portfolio via signed graphs and spectral clustering in the Brazilian stock market. Quality & Quantity. 2022 Aug;56(4):2325-40.
28. Zhen Y, Wang J. Community detection in general hypergraph via graph embedding. Journal of the American Statistical Association. 2023 Jul 3;118(543):1620-9.
29. Li M, Lu S, Zhang L, Zhang Y, Zhang B. A community detection method for social network based on community embedding. IEEE Transactions on Computational Social Systems. 2021 Feb 8;8(2):308-18.
30. Yazdanparast S, Havens TC, Jamalabdollahi M. Soft overlapping community detection in large-scale networks via fast fuzzy modularity maximization. IEEE Transactions on Fuzzy Systems. 2020 Mar 13;29(6):1533-43.
31. Zhuang D, Chang JM, Li M. DynaMo: Dynamic community detection by incrementally maximizing modularity. IEEE Transactions on Knowledge and Data Engineering. 2019 Nov 4;33(5):1934-45.