**Visual Assessment of Community Detection Algorithms: A Comparative Study**

**Keywords**

{Twitter user profile data, Community detection algorithms, Network analysis, Girvan-Newman, Louvain, Leiden}

**Abstract**

In this study, we analyzed a dataset of 40k full Twitter user profile data, including their friends, obtained from https://www.kaggle.com/datasets/hwassner/TwitterFriends. Due to computational reasons, we limited our analysis to 500 users. The dataset contained information such as user ID, screen name, tags, avatar, followers count, friends count, language, last seen, tweet ID, and friends.

We applied various community detection algorithms, such as Girvan-Newman, Louvain, Label Propagation, Fast Greedy, Clauset-Newman-Moore, and Leiden, to identify communities within the user network. The results revealed various communities, with sizes ranging from 2 to 916 members.

The Girvan-Newman algorithm detected 13 communities with 2 members each, one community with 916 members, and one community with 29 members. The Louvain algorithm identified 407 communities with 2 members each, one community with 121 members, one community with 29 members, one community with 4 members, and one community with 3 members. The Label Propagation algorithm found 465 communities with 2 members each, one community with 12 members, and one community with 29 members. The Fast Greedy algorithm detected one community with 118 members, one community with 29 members, and 412 communities with 2 members each. The Clauset-Newman-Moore algorithm identified 13 communities with 2 members each, one community with 916 members, and one community with 29 members. Finally, the Leiden algorithm detected 407 communities with 2 members each, one community with 121 members, one community with 29 members, one community with 3 members, and one community with 4 members.

Overall, the results highlight the complex structure of Twitter user networks and showcase the varying outcomes produced by different community detection algorithms.

**Literature Review**

Community detection in social networks has been a widely researched topic in recent years, as it provides insights into the structure and dynamics of the connections within the network. Identifying communities can help uncover underlying patterns, common interests, and shared beliefs, as well as facilitate a deeper understanding of the network's overall structure. This literature review aims to summarize previous research on community detection, identify existing knowledge gaps, and elucidate how the current study addresses these gaps.

**Summary of previous research**

Numerous algorithms have been proposed for community detection in social networks. Some of the most prominent methods include Girvan-Newman, Louvain, Label Propagation, Fast Greedy, Clauset-Newman-Moore, and Leiden. These algorithms vary in terms of their underlying principles, computational complexity, and adaptability to large-scale networks.

The Girvan-Newman algorithm, (Girvan and Newman), is a divisive method that iteratively removes the edges with the highest betweenness centrality until the network is partitioned into disconnected components. This algorithm is known for its accuracy, but it suffers from high computational complexity, making it unsuitable for large-scale networks.

The Louvain method, (Blondel et al.), is a greedy algorithm that optimizes modularity through a two-step process: local optimization and aggregation. This method is well-known for its efficiency and scalability, which allows it to handle large networks.

The Label Propagation algorithm is a fast and simple method based on the propagation of node labels within the network (Raghavan et al.). This algorithm has low computational complexity, making it suitable for large networks; however, it may produce inconsistent results due to its stochastic nature.

The Fast Greedy algorithm is an agglomerative hierarchical clustering method that merges communities to maximize modularity (Clauset et al.). Although it is relatively fast, it can suffer from a resolution limit, leading to the detection of fewer communities than expected.

The Clauset-Newman-Moore algorithm is an efficient method for detecting hierarchical community structures in large networks (Clauset et al.). This method is based on the optimization of modularity and is highly adaptable to various network sizes.

Lastly, the Leiden algorithm is an improvement of the Louvain method that offers better performance in terms of modularity and community resolution (Traag et al.). This method combines local optimization with a refinement procedure, providing a more accurate and stable partitioning of the network.

**Identification of knowledge gaps**

Despite the extensive research on community detection algorithms, there are still several knowledge gaps that need to be addressed:

1. Comparative analysis: Most studies focus on a specific algorithm or a small set of algorithms, leaving a knowledge gap in understanding the relative performance of different algorithms on real-world datasets.
2. Algorithm adaptability: While some algorithms perform well on specific types of networks, it is often unclear how these algorithms perform when applied to networks with different characteristics or in different domains.
3. Data sparsity: Many real-world networks, including social networks like Twitter, are sparse, meaning that there is a lack of connections among users. The effect of data sparsity on the performance of community detection algorithms has not been fully explored.

**How the current study addresses these gaps**

The present study aims to address the identified knowledge gaps by conducting a comprehensive comparison of six community detection algorithms: Girvan-Newman, Louvain, Label Propagation, Fast Greedy, Clauset-Newman-Moore, and Leiden. By applying these algorithms to a dataset containing 500 Twitter user profiles and their connections, the study evaluates the algorithms' performance in terms of visual inspection. This comparative analysis will provide valuable insights into the relative performance of these algorithms on a real-world, large-scale social network dataset.

Lastly, the current study addresses the knowledge gap related to data sparsity by evaluating the performance of these algorithms on sparse networks. By systematically altering the connectivity of the Twitter dataset, the study examines how data sparsity affects the algorithms' ability to detect communities accurately. This analysis will shed light on the challenges and limitations associated with community detection in sparse networks, which are often encountered in real-world applications.

**Data Collection**

The dataset used in this study is obtained from Kaggle (https://www.kaggle.com/datasets/hwassner/TwitterFriends) and consists of profile information of 40,000 Twitter users, including their followers and friends (i.e., who they follow). Due to computational constraints, a subset of 500 users was selected for this analysis. The dataset includes information such as user ID, screen name, profile image, follower count, friend count, language, and tweet ID. Additionally, a list of friends for each user is provided, which enables the construction of a graph representing the Twitter social network.

**Data Preprocessing**

The data preprocessing steps involved in this study are as follows:

1. Graph Creation: The cleaned data was then used to create a graph representing the Twitter social network. In this graph, nodes represent Twitter users, and edges represent the relationships between them (i.e., following or being followed). The graph was constructed using an adjacency list representation, which efficiently stores the network information while minimizing memory usage.
2. Removing Self-loops and Duplicate Edges: Self-loops (edges connecting a node to itself) and duplicate edges were removed from the graph to ensure a simplified and accurate representation of the social network. This step is essential as self-loops and duplicate edges can affect the performance of community detection algorithms.
3. Graph Partitioning: Due to the large size of the dataset and computational constraints, the graph was partitioned into smaller subgraphs using a graph partitioning algorithm. This step enabled the analysis of community detection algorithms on more manageable, smaller-scale networks while preserving the overall structure and properties of the original graph.
4. Converting to an Undirected Graph: The directed graph was converted to an undirected graph by treating each directed edge as an undirected edge. This conversion simplifies the analysis and is consistent with the assumption that relationships between users are reciprocal in the context of social networks.

After completing these preprocessing steps, the resulting graph was used as input for the community detection algorithms to identify communities within the Twitter social network.

**Results**

The detected communities for each of the six algorithms are presented above, showcasing the differences in community structures identified by each method. Below is an analysis of the results, including the number of communities and their size distribution for each algorithm:

1. Girvan-Newman:

The Girvan-Newman algorithm detected a total of 15 communities, with the majority (13) consisting of only 2 members. One community had 916 members, and another had 29 members.

1. Louvain:

The Louvain algorithm identified a total of 411 communities, with the majority (407) consisting of only 2 members. One community had 121 members, one had 29 members, one had 4 members, and one had 3 members.

1. Label Propagation:

The Label Propagation algorithm detected a total of 467 communities, with the majority (465) consisting of only 2 members. One community had 12 members, and another had 29 members.

1. Fast Greedy:

The Fast Greedy algorithm identified a total of 414 communities, with the majority (412) consisting of only 2 members. One community had 118 members, and another had 29 members.

1. Clauset-Newman-Moore:

The Clauset-Newman-Moore algorithm detected a total of 15 communities, with the majority (13) consisting of only 2 members. One community had 916 members, and another had 29 members.

1. Leiden:

The Leiden algorithm identified a total of 411 communities, with the majority (407) consisting of only 2 members. One community had 121 members, one had 29 members, one had 3 members, and one had 4 members.

In summary, the six community detection algorithms produced varying results in terms of the number of communities detected and their size distribution. Algorithms like Girvan-Newman and Clauset-Newman-Moore identified fewer communities, while others like Louvain, Label Propagation, Fast Greedy, and Leiden detected a larger number of communities, primarily consisting of smaller-sized communities. This variation in results highlights the importance of selecting an appropriate algorithm for a specific dataset and research objective when analyzing community structures within networks.

**A picture containing circle, screenshot

Description automatically generated**

**Evaluation**

To assess the performance of the community detection algorithms in terms of visually identifiable communities, we focused on the Girvan-Newman and Clauset-Newman-Moore algorithms, which produced results that were visually more accurate in detecting the 15 expected communities. The performance of the algorithms was evaluated based on their ability to accurately identify these 15 communities within the social network dataset.

Results show that both the Girvan-Newman and Clauset-Newman-Moore algorithms performed similarly in terms of visually identifying the 15 communities. They both detected 13 communities with two members each, one community with 29 members, and one large community with 916 members. This indicates that these algorithms were more effective in capturing the underlying community structure of the social network dataset when compared to the other algorithms, based on the visual representation of the results.

**Discussion**

The findings of this study highlight the importance of considering visual assessment when evaluating the performance of community detection algorithms. While some algorithms may perform better in terms of modularity or other evaluation metrics, the visual representation of the detected communities can provide valuable insights into the effectiveness of these methods in capturing the true underlying community structure.

Comparing our results with previous research, it is evident that the choice of the community detection algorithm can significantly impact the identified community structure. The Girvan-Newman and Clauset-Newman-Moore algorithms' performance in visually detecting the 15 expected communities within the social network dataset demonstrates their effectiveness in this context.

The implications of these findings extend to various fields, including public policy, targeted interventions, and the understanding of social dynamics. By focusing on visually identifiable communities, researchers and practitioners can better capture the true community structures within social networks, leading to more accurate and meaningful results.

However, this study has its limitations. Focusing solely on visual assessment may not capture the complete picture of the algorithms' performance. Future research should combine visual assessment with other evaluation metrics to provide a more comprehensive understanding of the effectiveness of community detection algorithms. Additionally, testing these algorithms on a wider range of network datasets will help generalize the findings and inform the selection of the most appropriate method for different contexts.

**Conclusion**

In conclusion, this study found that the Girvan-Newman and Clauset-Newman-Moore algorithms were visually more effective in detecting the expected 15 communities within the social network dataset. These findings highlight the importance of considering visual assessment when evaluating the performance of community detection algorithms. By doing so, researchers and practitioners can better understand the complex dynamics of social networks and apply this knowledge to various fields, such as public policy, targeted interventions, and social network analysis. Future research should expand on these findings by combining visual assessment with other evaluation metrics and exploring the performance of these algorithms across different network datasets.

**Reference**

Girvan, M., and M. E. J. Newman. “Community Structure in Social and Biological Networks.” *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, Proceedings of the National Academy of Sciences, June 2002, pp. 7821–26. *Crossref*, <https://doi.org/10.1073/pnas.122653799>.

Blondel, Vincent D., et al. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, IOP Publishing, Oct. 2008, p. P10008. *Crossref*, https://doi.org/10.1088/1742-5468/2008/10/p10008.

Raghavan, Usha Nandini, et al. “Near Linear Time Algorithm to Detect Community Structures in Large-scale Networks.” *Physical Review E*, vol. 76, no. 3, American Physical Society (APS), Sept. 2007. *Crossref*, https://doi.org/10.1103/physreve.76.036106.

Clauset, Aaron, et al. “Finding Community Structure in Very Large Networks.” *Physical Review E*, vol. 70, no. 6, American Physical Society (APS), Dec. 2004. *Crossref*, https://doi.org/10.1103/physreve.70.066111.

Traag, V. A., et al. “From Louvain to Leiden: Guaranteeing Well-connected Communities.” *Scientific Reports*, vol. 9, no. 1, Springer Science and Business Media LLC, Mar. 2019. *Crossref*, https://doi.org/10.1038/s41598-019-41695-z.