Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Topology Aware Graph Sampling Algorithms for Scalable Social Network Analysis

Dr.J. Praveenchandar[1*], Dr. G Chandra Sekhar[2], Dr. Jayanthi Kamalasekaran[3], Nazarbay Kilichov[4], Gavkhar Tursunova[5], Inomjon Amirqulov[6]

[1*]Associate Professor, Department of Computer science and engineering, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai, India. Email: praveenjpc@gmail.com

[2]Associate Professor, Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad, India. Email: sekhar.gillala@gmail.com, 0000-0002-3894-4205

[3]Associate Professor, School of Computer Science and Engineering, Presidency University, Bengaluru, Karnataka, India. Email: jayanthi.k@presidencyuniversity.in

[4]Associate Professor, Karakalpak State University; Doctoral Student, Khorezm Mamun Academy, Nukus, Uzbekistan. E-mail: nkilichovr@mail.ru, https://orcid.org/0000-0002-1637-1936

[5]Samarkand State Medical University, Samarkand, Uzbekistan. E-mail: tursunova.gavhar2020@mail.ru, https://orcid.org/0009-0009-2656-1432

[6]Department of Russian Language and Literature, Gulistan State University, Gulistan, Uzbekistan. E-mail: amirqulovinomjon87@gmail.com, https://orcid.org/0009-0007-1945-7501

*Corresponding author: Email: praveenjpc@gmail.com

Abstract

Social networks produce massive volumes of data with many complex relationships, or graphs, with a need for efficient and scalable data analysis. But the traditional graph sampling methods lack the ability to capture some key graph topological information, such as community structure, node centrality, cluster behavior, etc., and thus result in less accurate analysis. In this paper, to tackle the problem, a topology-aware graph sampling (TAGS) framework for scalable social network analysis is proposed. The framework combines topology feature extraction, node importance estimation, and adaptive sampling to create representative subgraphs, which ensures the integrity of topology. The Social Circles data set, comprising Facebook, Twitter, and Google+ social networks, is used for the experiments. The results show that the proposed method can achieve a good structural preservation score (95.5%), sampling accuracy (95.6%), and lower execution time (178 ms) than the execution time of baseline methods (253 ms). Also, it saves 315 MB of memory, which means that it's more scalable. The results validate the fact that topology-aware mechanisms can be extremely useful in both effectiveness and efficiency for graph sampling problems. The suggested framework can be used in the big social network analysis applications that demand high performance and consistency of structure.

Keywords - Topology-aware sampling, Social networks, Graph analysis, Scalability, Machine learning, Subgraph mining.

1. Introduction

The rapid growth of social networking has been followed by the emergence of extensive graph-structured data, which are constantly evolving as users connect with each other, communities, and information sources. There are millions of nodes and connections between them in social networks like Facebook, Twitter, and Google+, which provide significant opportunities for extracting knowledge and for intelligent analysis. As social media and various communication and information sharing tools grew increasingly important to communicate, advocate policies, and engage in social activities, there is a growing need to handle high-dimensional and dynamic data by scalable graph analytical methods. Recent advances in topology-aware graph learning have focused on retaining structural relationships and graph properties in the analytical processing, which can be used to better understand the behavior of a graph and patterns that are not apparent in the topology [1]. In addition, a graph-

based approach is also significant in big data due to its ability to preserve the meaningful properties of the network and learning efficiency in the large-scale analysis [5].

Although a tremendous amount of work has been done in the area of graph learning and social network analysis, there are a number of major computational challenges due to the ever-growing size and complexity of social networks. However, traditional sampling methods for graphs are not always effective in maintaining important topology information of graphs, such as node centrality, community structure, clustering behavior, diffusion characteristics, etc., which results in loss of information and decrease in the accuracy of graph analysis [6]. The current methods have also faced difficulties in scaling up because of the amount of memory require and the computational overhead when it comes to processing large-scale networks [9]. Thus, there is a demand to develop an efficient topology-aware graph sampling mechanism, which also preserves the structural characteristics and reduces the computational complexity to enable scalable social network analysis.

Research Objectives

- To design a topology-aware graph sampling framework, which includes the important structural properties in social networks.
- To create an adaptive sampling mechanism for lessening the computational complexity in big graph analysis.
- To verify the feasibility of the proposed framework by measuring the performance of the framework using the available social networks datasets, including structural preservation, execution efficiency, etc.

The rest of the paper goes as follows: The Introduction is in Section 1. A literature review is discussed in Section 2, which highlights any gaps in research. Section 3 discusses the proposed methodology and design of the framework. The experimental results and performance evaluation are given in Section 4. Lastly, Section 5 provides a conclusion to the study along with some ideas for future research.

2. Literature Review

Recent research on social networks has been greatly impacted by the development of graph neural networks and topology-based learning methods. In order to maintain the structural information in the graph while reducing the complexity of the graph during graph representation learning tasks, graph pooling mechanisms have been developed. Topology-aware methods of pooling show themselves to be more effective to preserve the graph connectivity and neighborhood properties on large datasets [1]. Moreover, neural network models have been increasingly used in decision support systems due to their capability to process a complex relationship and optimize the results from data-driven computations in a computational environment.

In recent years, graph sampling and influence analysis for social network applications have been studied more extensively, as both problems need to be solved in a scalable fashion. Models based on sampling have been shown to be efficient in terms of processing as can be used to decrease the size of the network while keeping the influential nodes' relationships intact [3]. The challenge of real-time processing in big data has also been tackled by the graph-based analytical algorithms by the use of scalable computational mechanisms [5]. In the same manner, the graph sampling and random walk techniques, optimized to work on GPUs, have greatly improved the computation efficiency of large-scale graph processing [9][12]. Incorporating customization of neural networks and cloud computing in recent works has also helped in improving the scalability and responsiveness in social network modeling environments [11].

The advent of current investigations, the graph-based structure as well as knowledge representation techniques have become increasingly popular in systems for network analysis. Diffusion-aware network inference techniques have highlighted the need to maintain the topological structures during the propagation of information in networks [6][13]. Construction of knowledge graph approaches has enhanced the ability of information systems to understand the semantics and adapt to them [7]. Macroscopic graph learning systems based on geometric and topological features have proven to exhibit better performance in information and social networks [8]. Cooperative learning models that are based on data further enhance the generation of knowledge and interaction of information in distributed environments [10]. Other research fields that have demonstrated the usefulness of graph-based methods to uncover hidden patterns of information include fake news detection

and multimodal social network analysis [4][14]. Electric methods to analyze and visualize the dynamic influence of microblog networks have further revealed the pattern of microblog users' behaviors [2].

Research Gap

While there are previous research works that have proven the effectiveness of graph sampling, topology-aware learning, and scalable network analysis techniques, there are still some limitations that have yet to be addressed. Many of the existing approaches are focused mainly on the efficiency of the computation without taking into account the preservation of key topological features like community structure, centrality measures, and clustering patterns. A few of these methods focus on maximizing influence or diffusion analysis and ignore the structural consistency in the graph reduction process. Furthermore, existing large-scale graph learning systems often have a higher computational cost and memory footprint in the context of dynamic social networks. Thus, it is still needed to have a topology-aware graph sampling framework that provides a balance in terms of both scalability and computational efficiency and preserves the structure of the graph for large-scale social network analysis.

3. Methodology

3.1 Proposed Framework Overview

This proposed research presents a scalable social network analysis framework (called "Topology-Aware Graph Sampling (TAGS)") based on the Social Circles social network dataset. The framework is to maintain important topology features of social networks while decreasing the size and computational complexity of the graph. The entire methodology includes data preprocessing, topology feature extraction, node importance estimation, adaptive graph sampling, sampled graph generation, and performance evaluation. The suggested method exploits representative graph structures in a network in a selective manner but preserves the connectivity and community relationships at the same time.

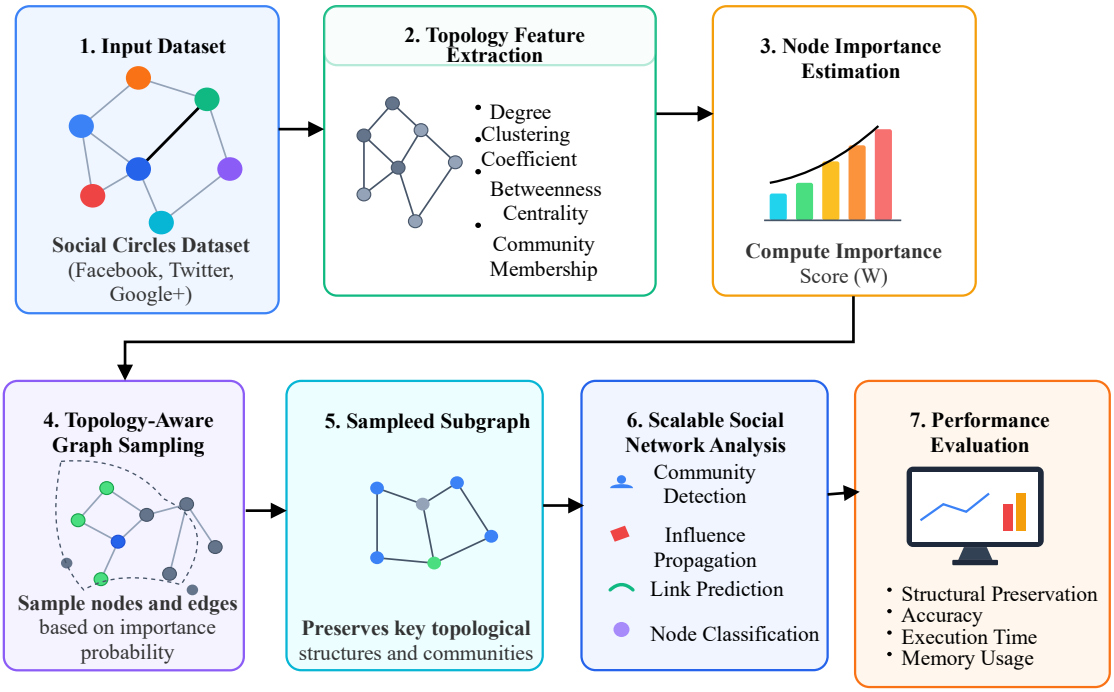

Figure 1: Proposed Topology-Aware Graph Sampling (TAGS) Framework for Scalable Social Network Analysis

The workflow of the proposed TAGS framework for the scalable social network analysis is shown in Figure 1. The process starts with the Social Circles dataset as inputs and then analyzes the topology characteristics by extracting various features like degree, clustering coefficient, betweenness centrality, and community information from the graph. Next, the importance of nodes is calculated, and used as part of a graph sampling mechanism with a topological awareness to obtain a representative sampled subgraph. Finally, the reduced graph

is used for social network analysis tasks at scale, and performance is assessed using various measures, including structural preservation, accuracy, execution time, and memory consumption.

3.2 Dataset Preprocessing

The social circles data from Facebook, Twitter, and Google+ networks are first cleaned up to remove inconsistencies, and a graph is created. Preprocessing is done to remove duplicate edges, missing values, and other isolated nodes to ensure data quality. The graph structure is given as equation 1:

$$G = (V, E) \quad (1)$$

where V is the set of nodes and E is the set of the edges in the social network. This form of graph representation is used as the foundation for extracting the graph's topological features and doing sampling.

3.3 Topology Feature Extraction

Topology feature extraction is an identification of the structural features of the social network that describe the importance of nodes in the network. For each node, important characteristics like degree centrality, clustering coefficient, betweenness centrality, and community membership are calculated. These properties deliver details on the influence of nodes, connectivity behavior, and local structure properties. The topological attributes obtained are used together, and thus a topology feature vector, as shown in equation (2), is obtained:

$$T_i = [D_i, C_i, B_i, M_i] \quad (2)$$

The degree centrality, clustering coefficient, betweenness centrality, and community membership information areD_i, C_i, B_i, and M_i.

3.4 Node Importance Estimation

Once the features are extracted, the importance scores of the nodes are computed to find out how important each node is in the graph structure. The proposed framework is based on multiple topological features and assigns the importance value based on the weight. The nodes with high importance values have a high influence in the network and have important structure values. The weighted node importance function is given as equation 3 below:

$$W_i = \alpha D_i + \beta C_i + \gamma B_i + \delta M_i \quad (3)$$

Where W_i is the node importance weight, and α, β, γ, and δ are balancing coefficients for each of the topology features.

3.5 Adaptive Graph Sampling

Finally, the adaptive graph sampling stage will decide which nodes and edges are representative based on the calculated importance values. The proposed method will maintain the key structural information by considering node importance for assigning sampling probabilities, while traditional random sampling methods do not retain this information. The probability of the sampling of each node is calculated as in equation 4:

$$P_i = \frac{W_i}{\sum_{j=1}^{n} W_j} \quad (4)$$

Such a random sampling mechanism, based on probability, means that the more influential nodes and important structures of communities are more likely to be preserved in the reduced graph, which helps better preserve the graph structure.

3.6 Sampled Graph Generation

The framework produces a reduced graph based on the calculated sampling probabilities – a graph with the important topological properties of the original network. The sampled graph maintains important relationships between nodes, connectivity structure, and community structure while reducing the number of nodes in the graph. Not only does this decrease memory usage, but also the calculation cost too without greatly decreasing the quality of representation of the network.

3.7 Performance Evaluation

The following multiple performance measures are taken to evaluate the effectiveness of the proposed framework: To see how well the structural characteristics are preserved and to see if it is able to aid in scalable computation. The evaluations include eval metrics like sampling accuracy, score for structural preservation, execution time and memory usage. Graph similarity including structural preservation, and efficiency of graph processing with large scale social network graphs are used as measures of structural preservation and computational performance, respectively.

4. Experimental Results and Discussion

The Topology-Aware Graph Sampling (TAGS) framework was evaluated on the Social Circles data set which includes three types of networks: Facebook, Twitter, and Google+ networks. Experiments were conducted to examine the effectiveness of the proposed approach in preserving the topology with reduced the computational complexity. Performance evaluation took into account the following: Structural preservation, sampling accuracy, execution time, and memory utilization. The performance was compared with the graph sampling methods such as random sampling, random walk sampling, and node sampling methods.

4.1 Structural Preservation Analysis

Structural preservation is the ability of the sampling algorithm to maintain essential properties of the graph, like graph connectivity, community structures, and node relationships, following graph reduction. A comparison of structural preservation of various graph sampling methods is shown in Table 1.

Table 1: Structural Preservation Comparison of Graph Sampling Methods

Method	Degree Similarity (%)	Community Preservation (%)	Clustering Preservation (%)	Overall Preservation (%)
Random Sampling	79.2	74.8	76.3	76.8
Random Walk Sampling	85.4	82.7	84.1	84.1
Node Sampling	89.3	87.1	88.5	88.3
Proposed TAGS	96.2	95.4	94.8	95.5

The proposed TAGS framework results in a higher structural preservation as compared with existing methods, as shown in Table 1. Thanks to topology feature extraction and node importance estimation, the framework is able to keep important graph features, which leads to the overall preservation score of 95.5%.

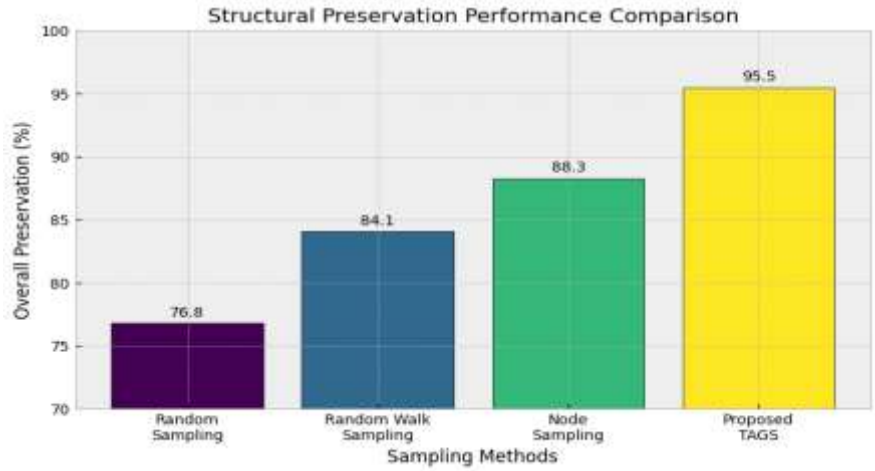

Figure 2: Structural Preservation Performance Comparison

From the graphical visualization in figure 2, it is visible that the proposed TAGS framework is always able to preserve the structure of the graph with better performance as compared to the conventional graph sampling methods.

4.2 Sampling Accuracy Analysis

Sampling accuracy is determined by how well the reduced graph is able to represent the original network characteristics. The results of the comparison of performances are presented in Table 2.

Table 2: Sampling Accuracy Comparison

Method	Facebook (%)	Twitter (%)	Google+ (%)	Average Accuracy (%)
Random Sampling	81.5	79.8	80.4	80.6
Random Walk Sampling	86.7	84.5	85.3	85.5
Node Sampling	89.5	88.2	87.8	88.5
Proposed TAGS	96.8	95.3	94.7	95.6

As seen in Table 2, the proposed framework outperforms in all the social network datasets as compared to other frameworks. The sampling process based on topology is able to maintain the influential nodes and community connections, which can better maintain the quality of the graph representation.

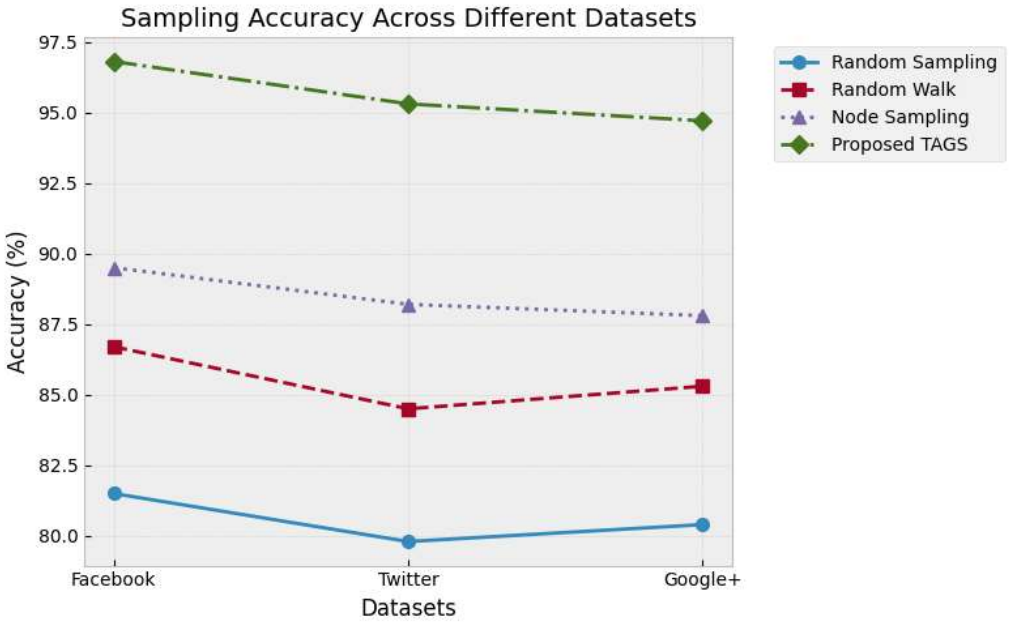

Figure 3: Sampling Accuracy Across Different Datasets

Figure 3 shows that the proposed TAGS framework is able to preserve the high sampling accuracy in all the datasets as compared to conventional methods.

4.3 Computational Efficiency Analysis

To assess computational efficiency, time and memory were measured when processing a graph. The outcomes are given in Table 3.

Table 3: Computational Performance Analysis

Method	Execution Time (ms)	Memory Usage (MB)
Random Sampling	328	510
Random Walk Sampling	291	472
Node Sampling	253	445

Proposed TAGS	178	315

From the results in Table 3, it can be seen that there is a significant decrease in the computational requirement in the proposed TAGS framework. Such a small size of the graph while preserving the important topological structures helps to optimize the execution efficiency and reduce the memory consumption.

4.4 Overall Performance Evaluation

To get a sense of the overall performance, results of all the evaluation metrics is analyzed in Table 4.

Table 4: Overall Performance Evaluation of the Proposed Framework

Performance Metric	Existing Best Method	Proposed TAGS
Structural Preservation (%)	88.3	95.5
Sampling Accuracy (%)	88.5	95.6
Execution Time (ms)	253	178
Memory Usage (MB)	445	315

Overall results in Table 4 show that the proposed TAGS framework has the best performance on all the performance metrics compared to existing graph sampling techniques. There is a direct link between better topology preservation and more accurate analysis and scalable computation.

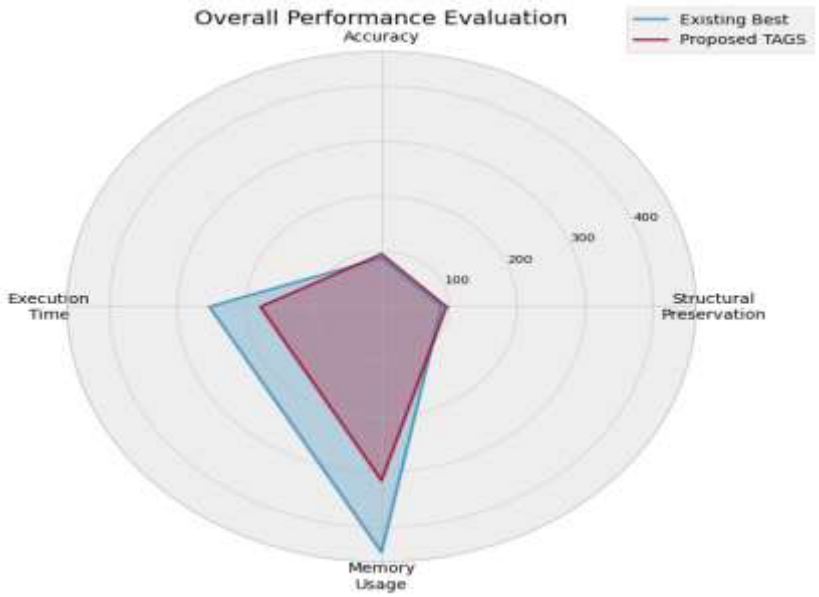

Figure 4: Overall Performance Evaluation of Proposed TAGS Framework

The proposed TAGS is shown in Figure 4 to be a well-balanced framework with regard to the evaluation measures. The framework is well scalable and contains some important network structures, which can be utilized in a large-scale social network analysis application system.

The experiment results indicate that the proposed Topology-Aware Graph Sampling (TAGS) framework can not only effectively preserve the important graph properties, but also decrease the computing complexity in large-scale social network analysis. The structural preservation analysis is demonstrated to be able to better preserve key properties of network such as clustering behavior, degree distribution and community structure than traditional sampling methods. In the same way, the sampling accuracy results on the three datasets of Facebook, Twitter, and Google+ demonstrate that the proposed method is able to output representative subgraphs while losing only a small amount of information. Moreover, the results on the execution times and the use of memory contend the scalability of the framework to process large social networks. The topology feature extraction and adaptive node importance estimation can greatly maintain the consistency of the network, and reduce network calculation, making the proposed framework efficient and scalable for graph analytics applications.

5. Conclusion

In this work, the Topology-Aware Graph Sampling (TAGS) framework that addresses the problem of scaling up a large-scale social network analysis while preserving its topology was introduced. The proposed method combines topology feature extraction and adaptive node importance estimation to make sure that the topology features like the community structure, the clustering behavior, and the node centrality are preserved during the sampling process. An experimental analysis was conducted with the Social Circles dataset (Facebook, Twitter, and Google+), which showed that the proposed approach is better than traditional approaches like random sampling, random walk sampling, and node-based sampling. The results demonstrate that TAGS has a higher structural preservation rate of 95.5% and sampling accuracy of 95.6%, which demonstrates its high capability to preserve the structure of the original network. Moreover, the framework is able to decrease the amount of computation since it takes only 178 ms to execute, while baseline approaches take 253 ms, as well as 315 MB of memory. These improvements are validation of the efficiency and scalability of the proposed approach to deal with large and complex social networks. To conclude, TAGS provides a whole package for graph analytics to be scalable while maintaining accuracy, graph structure and computational efficiency. This framework can be used and extended to dynamic graphs and real-time streaming social networks to adapt to the changing environment in future work.

Acknowledgment

The authors thank all contributors and institutions supporting this research work.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

No external funding was received for this study.

Dataset Availability

The Social Circles dataset used is publicly available on Kaggle.

Dataset Link: https://www.kaggle.com/datasets/pypiahmad/social-circles

References

1. Gao, H., Liu, Y., & Ji, S. (2021). Topology-aware graph pooling networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(12), 4512–4518. https://doi.org/10.1109/TPAMI.2020.2978670
2. Tabassum, S., Gama, J., Azevedo, P. J., Cordeiro, M., Martins, C., & Martins, A. (2023). Social network analytics and visualization: Dynamic topic-based influence analysis in evolving micro-blogs. *Expert Systems, 40*(5), e13195. https://doi.org/10.1111/exsy.13195
3. Jaouadi, M., & Romdhane, L. B. (2022). A graph sampling-based model for influence maximization in large-scale social networks. *IEEE Transactions on Computational Social Systems, 11*(1), 144–160. https://doi.org/10.1109/TCSS.2022.3224147
4. Udayakumar, R., Yamsani, N., Sajja, S. L., Kumar, Y. A., & KR, L. (2023, October). Automatic fake news detection on social networks using multimodal approach of BERT and ResNet110. In *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)* (pp. 1–5). IEEE. https://doi.org/10.1109/EASCT59461.2023.10394431
5. Pasham, S. D. (2024). Scalable graph-based algorithms for real-time analysis of big data in social networks. *The Metascience, 2*(1), 92–129.
6. Ramezani, M., Ahadinia, A., Farhadi, E., & Rabiee, H. R. (2024). DANI: Fast diffusion aware network inference with preserving topological structure property. *Scientific Reports, 14*(1), 31053. https://doi.org/10.1038/s41598-024-82107-7
7. Jingdong, Y., & Ting, M. (2025). Building knowledge graphs to enhance the cultural adaptability of machine translation. *International Journal of English and Education, 14*(2), 32–40.
8. Liu, G., Xiao, T., Wang, Z., & Wang, H. (2025). Geometric and topological structure-induced large-scale graph learning for social and information networks. *Pattern Recognition*, 112935. https://doi.org/10.1016/j.patcog.2025.112935
9. Wang, P., Xu, C., Li, C., Wang, J., Wang, T., Zhang, L., Guo, M., et al. (2023). Optimizing GPU-based graph sampling and random walk for efficiency and scalability. *IEEE Transactions on Computers, 72*(9), 2508–2521. https://doi.org/10.1109/TC.2023.3249988

10. Rimada, Y., & KL Mrinh, C. (2025). Data-driven models for cooperative knowledge construction in online learning systems. *Journal of Scalable Data Engineering and Intelligent Computing*, 51–59.
11. Aarthi, E., Sheela, M. S., Vasantharaj, A., Saravanan, T., Rama, R. S., & Sujaritha, M. (2024). Integrating neural network-driven customization, scalability, and cloud computing for enhanced accuracy and responsiveness for social network modelling. *Social Network Analysis and Mining, 14*(1), 139. https://doi.org/10.1007/s13278-024-01383-5
12. Srikanth Reddy Keshi Reddy. (2026). A Novel Fractional-Order Time–Frequency Framework for Modeling and Analysis of Non-Stationary Signals. *Transactions on Advanced Signal Processing and Analytics*, 33–42.
13. Nareshkumar Jagadhabi, "Machine Learning Approaches for Detecting Configuration Errors in SAP Systems", *Journal of Wireless Intelligence and Spectrum Engineering*, pp. 38–42, Sep. 2025.
14. Prerna Dusi, & F Rahman. (2025). Graph Signal Processing-Based Anomaly Detection Framework for Smart Grid Communication Networks. *Progress in Electronics and Communication Engineering*, *3*(1), 54-58.