Anonymous Social Network Analysis with PageRank

ABSTRACT

The PageRank algorithm is one of the link analysis algorithm which is commonly used in graph structured data to rank the nodes which shows their importance in the graph data.It is one of the success factors behind today's popular web search engines including Google, and Yahoo etc., with Google on the top since long.However the growth in number of social media users continues to increase,traditional approaches are no longer supported for finding influential nodes as it leads unsolvable.To,address this challenge,In this paper,we have proposed a link analysis of nodes in the social network using PageRank algorithm.The approach involves three phases: Pre-process social network data,PageRank score calculation with Graph Construction and Influence Identification.We also integrate with Logistic Regression to enhance the accuracy of our predictions..By Implementing this approach we are able to find influential nodes efficiently in the large social network datasets and make accurate about few scenarios such as Network Health,Viral Marketing,Opinion and Trend Shaping,etc.We anticipate that this approach will be particularly useful in the context of finding influential nodes in large social network datasets which can be processed quickly and accurately.

INTRODUCTION

Data Mining

The process of extracting information to identify patterns,trends and useful data that would allow us to take data-driven decisions from huge sets of data is called data mining.In other words we can say that Data mining is the process of investigating hidden patterns.It is primarily used to convert raw data into useful information. It encompasses a range of methodologies and techniques from statistics, machine learning and database management.Data mining has applications in various fields, including business, healthcare, finance, marketing, scientific research, and more.Data mining techniques are typically applied to structured data, which is organized in tables with rows and columns, like databases and spreadsheets.

It is also known as Knowledge Discovery in Database(KDD).Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation are just a few of the crucial elements that make up the Knowledge Discovery in Databases (KDD) process.We have access to a number of techniques in the field of data mining, including C4.5, K-Means, Apriori, Support Vector Machines, and PageRank.Clustering, classification, regression, sequential patterns, and association rules are just a few of the major subcategories that these data mining technologies fall under.

The various types of services in data mining are Text Mining,Web Mining,Audio Mining,Video Mininga,Pictorial data Mining,Social Media Mining.The primary goal of data mining is to uncover hidden patterns and relationships within data that can aid in decision-making, prediction, and understanding complex phenomena.

Graph Mining

Graph mining focuses on analyzing data structured as graphs. A graph is a collection of nodes (entities) connected by edges (relationships or interactions). Graphs are used to represent relationships between entities, and they can be directed (edges have a specific direction) or undirected (edges have no direction). Graph mining involves extracting patterns, structures, and insights from these graphs. It can involve tasks like community detection, centrality analysis, subgraph pattern mining, and more.Graph mining is a multidisciplinary field that combines concepts from graph theory, machine learning, data mining, and network science. It has applications in a wide range of domains, including social media analysis, biology, transportation, recommendation systems, fraud detection, and more.

Graph mining encompasses a range of algorithms and techniques. The graph mining algorithms include graph traversal methods like BFS and DFS, centrality analysis, community detection, and subgraph pattern mining.Graph mining techniques encompass the application of the PageRank Algorithm, graph classification and clustering, graph matching and similarity analysis, temporal graph mining, and graph partitioning for various data analysis tasks.The application of graph mining includes Social Network Analysis,Recommendation Systems,Transportation and Infrastructure,Fraud Detection and Security,Epidemiology and Disease Spread Analysis,Social Media and Online Behavior Analysis,Web Graph Analysis.The main objective of your project is to leverage graph mining techniques

to analyze the structure and relationships within a web graph. This involves extracting valuable insights from the interconnected web pages to understand patterns, influence, and connectivity.

Data Mining Vs Graph Mining

Graph mining can be considered a specialized form of data mining that focuses specifically on the representation of data as images. In other words, visual search is a search for information that considers data sets. Many techniques used in traditional literature can be used for graph mining by identifying nodes, edges, and their attributes.

Data mining is generally applied to structured and unstructured data, including text and numbers. Graph Explorer specializes in handling structured data in graphs; This makes it ideal for situations where relationships and connections are important. Additionally, graph mining can provide unique information that is difficult to find with traditional data mining techniques. For example, conversation analysis using Graph Explorer can reveal social patterns, interaction patterns, and important people that might be missed when looking at the same data in a paper chart. Each method has its own methods and algorithms and uses data based on underlying data and objectives.

In summary, where data search and visual search indicate the goal of extracting important information from large data sets, visual search specifically controls the data representation as images and focuses on the relationship between organizations. Both serve the general purpose of uncovering insights from data, and their suitability depends on the nature of the data and the specific goals of the analysis.

LITERATURE REVIEW

The literature review provides a summary of research papers that have implemented Graph Mining.The summaries focus on various aspects.In [J1],they have introduced **MIGDAC** graph mining algorithm which when compared to conventional methods,dramatically increases classification accuracy by using hierarchical graph representations and interestingness measures to find class-specific patterns.It uses graph mining techniques like FSG or gSpan, frequent to extract subgraphs.In [J2],they have used **GRADOOP** algorithm, a system for complex data analytics with graph mining capabilities which involves subgraph mining.It is used as a flexible data analytics solution with graph mining capabilities for complicated jobs. It displays its capabilities in a business intelligence setting, highlighting frequent subgraph mining in particular.In [J3], the focus is on capturing entities and relationships between them from multiple data sources and supporting both dense and frequent substructure discovery.The framework consists of five key modules: Graph Preprocessing, Graph Database, Frequent Substructure Discovery, Dense Substructure Extraction, and Graph Visualization. The framework aims to integrate data from various sources into a graph database, perform efficient **substructure discovery,** and provide interactive graph visualization for knowledge exploration.In [J4],it offers a generic framework for fault prediction using maintenance and alarm records, relationship mining, and Graph Neural Networks to forecast defects in a variety of mechanical equipment, providing a flexible method for equipment health management.The method entails organizing links between alarms and constructing an Alarm Track Graph (ATG) from alarm information. These graphs are analyzed using **Graph Convolutional Networks (GCN)** to forecast mechanical equipment problems. In [J5], they analyzed huge graph data utilizing the **Bulk Synchronous Parallel** (BSP) computing model, BPGM is a parallel data mining solution.The constraints of BPGM may be addressed, and its capabilities for huge graph data processing may be further improved, in future development.The Bulk Synchronous Parallel (BSP) paradigm is used by the BPGM system to do parallel data mining on huge graph datasets.In [J6], the method is finding common sub-community graphs among n-community graphs, with a focus on a case involving village communities. It successfully detects sub-community graphs using graph mining algorithms and adjacency matrices for comparison.Graph mining techniques and adjacency matrices are used in the methodology for finding **frequent sub-community graphs** within a collection of n-community graphs.Finally, real-world examples from the village communities are shown, demonstrating the successful recognition of these common sub-community graphs and demonstrating the usefulness of the approach.In [J7], they explained a graph mining approach for extracting sub-graphs, weak and strong links, and call patterns from telephone network data, illustrating how it can be used to research human behavior.The telephone network data is initially shown as a graph with people acting as nodes and

call interactions acting as edges.The programme then locates sub-graphs, highlighting teams of people who frequently communicate with one another.In [J8], they have outlined an approach for leveraging DNS data and graph mining to find rogue domains. In the beginning, DNS data is gathered to guarantee its accuracy and diversity. A DNS graph is then created, with domains acting as nodes and interactions as edges. In order to determine reputation scores for each node in the DNS graph, a modified **Belief Propagation** technique is then used.This method provides useful information about the field of malicious domain detection.In [J9], they explained an approach for building a dynamic social network from Facebook data that emphasizes user connections and makes use of pre-established principles to forecast users' upcoming social interactions. It emphasizes linkages between users and communities that aren't directly connected, making predictions based on things like friendship proximity and community user activity. The study emphasizes the importance of graph visualization of data for better comprehension and offers applications in marketing and the fusion of data mining and machine learning.The methodology for constructing a **dynamic social graph** from Facebook data involves several steps.In [J10], it is based on graph theory, a novel method for mining **Maximal Frequent Itemsets (MFI)** is suggested.This method overcomes a number of MFI mining-related concerns, such as high memory utilization, temporal complexity, partial findings, and repeated database searches.This strategy assures that all MFI are identified while requiring only one database scan in comparison to other approaches.It also uses less memory.

In [V1],They have performed Fake news detection based on Transformer Architecture which has two parts: encoder: to learn useful representations from the fake news data and decoder that predicts future behavior basing on observations from the past,a Graph Mining Approach used in Natural Language Processing.In [V2],They have Located Solar Panels in the locality using graph mining since the grid is in graph form.They found an algorithm that identifies graphical topology with time series data and then converted into gra[h-type data.Then the graph data is inserted into a Graph Neural Network and have done classification..In [V3],Typically different approaches of graph mining are used in different applications of graph mining so to bridge the gap,they are using algorithms like Pagerank to graph neural networks to create an algorithm which determines the type of algorithm to be used for the given data.In [V4],They have made an efficient machine based on a graph mining system which treats disks as an extension of memory.It treats intermediate data as tensor and implements half memory half disk storage for storing data,adopting an isomorphism checking strategy which uses an Eigenvalue based algorithm and tree isomorphism..In [V5],They understood behavior analysis of human interactions when it takes place.Graph mining is an efficient way to analyze and get required information.Since the data involved is dynamic in nature which changes with respect to time dynamic type of data,dynamic graph mining is used..In [V6],They have used communities as subsets of nodes in networks and represented them using graphs with different parameters and by improving one of the parameters better communities can be detected.They have used different approaches like graph partitioning and clustering..In [V7],they are using data of cyber attacks to identify malicious domains.Since the domain’s IP addresses change frequently to escape they are using the data in graph forms for accuracy.In[V8],They are detecting a web security vulnerability named Cross Site Request Forgery using graph mining techniques.They have used graph modeling,Finite State Machine analysis,parameter categorization and vulnerability testing to detect CSRF vulnerabilities..In [V9],They are extracting the graphical structures and features of panels and then separate them into different categories using attributed Region Adjacency Graphs.Then some frequent subgraphs are obtained using Frequent Subgraph MIning techniques.And for content based image retrieval purpose they are analyzed by isomorphism between the graphs and sort them.In [V10],They applied a graph mining system to sales data to understand consumer behavior.They represent the customer purchase behavior by a directed graph retaining temporal information and apply a graph mining technique to analyze the frequent occurring patterns.They use an example of cooking oil sales to analyze.

In [G1], Fault prediction has been introduced in equipment of health management by analyzing the relationship between alarm signals and fault events using **Apriori** algorithm.In [G2],Graph based representation of Customer Reviews for online stores has been introduced that transforms dataset containing free text reviews for online stores in graph-based form .In [G3], Graph mining Framework for finding and visualizing substructures using

graph database is presented by proposing a graph mining framework that captures entities and relations between entities from Different data sources and this framework further models this data as a graph and facilitates the substructure extraction.In [G4], Fake News Detection A Graph mining Approach is introduced to identify fake news stories using graph mining techniques.It was tested from real world data set and those experiments stated that it performs similarly to existing fake news detection algorithms.In [G5],Graph Mining for Classifying and Localizing Solar Panels in Distribution Grids is presented.Which is used to acquire approximate locations of solar users by Graph mining approach for solar panel localization.Here they used solar panel classification algorithm that identifies graphical topology with time series data.Based on this time series information they designed a graph construction algorithm and converted time series data to graph type data.In [G6],An Efficient way to Find Frequent Patterns Using Graph Mining and Network Analysis Techniques on United States Airports Network they used transportation network of airports in United States of America and applied graph data mining techniques and network analysis techniques on US airports and flight Data Sets.The paper presents valuable insights into how graph mining and network analysis techniques can be applied to analyze data related to U.S. airports and flights, helping to uncover frequent patterns and essential information for improving services and safety in the aviation industry.In [G7] Constraint-Based Graph Mining in Large Database focuses on systematically incorporating various constraints into the graph mining process to efficiently extract meaningful patterns from large databases.Provides insights into the challenges and methods of integrating constraints into the graph mining process, making it valuable for researchers and practitioners interested in data mining and graph analysis.In G[8],Topic Mining Based on **Graph Local Clustering** presents an approach for discovering thematically related document groups like topic mining in large document collections using graph local clustering.It focuses on using structure like hyperlinks in web collections to perform topic mining, taking advantage of language-independent and immune properties.It demonstrates the effectiveness of this approach with experiments on Wikipedia data. This paper can be valuable for researchers interested in document clustering and topic mining in large-scale web collections.In G[9],Consumer Behavior Analysis by Graph Mining Technique discusses the application of graph mining to understand consumer behavior based on sales transaction data.Introduces the application of graph mining to consumer behavior analysis, particularly in the context of health-related product purchases. It highlights the potential of graph mining techniques to uncover meaningful patterns in complex purchasing behavior, and the authors suggest that this approach can be applied to the business and marketing domain.In [G10],they have discussed the application of graph mining, particularly the **SUBDUE** algorithm, to unsupervised image analysis, focusing on background subtraction from videos.The aim is to apply graph mining techniques to unsupervised image analysis, where knowledge can be discovered without prior supervision.This paper presents an innovative application of graph mining, with a specific focus on using the SUBDUE algorithm, for the purpose of background subtraction in videos that are recorded by surveillance cameras.

In [K1], they have introduced **Random walk** based graph mining approach where a random sequence of points are selected on the graph,which address the data needs for the client.The Random Walk is implemented by constructing heterogeneous graph by integrating information from multiple resources followed by that movement in the graph will be starts from some vertex and at each time moves to other step.In [K2], they presented an extension of the previous algorithm(Belief-Propagation) which doesn't cover all the components of the graph. So, they have introduced a law of total probability based on the relationship between domains and clients to calculate the malicious score which has great accuracy.The **Belief-Propagation** algorithm is also known as sum-product message passing algorithm which is designed to transfer information between objects on an undirected graph obeying Markov properties(The next state depends upon only on the present state but not the past state).In [K3], they have utilized sub-graph or patterns from weighted graphs with edge weight detection rather node weight Discovering **subgraphs or patterns** appearing frequently in a set of graphs.In [K4], they have partitioned the large community graph into small community graphs.The knowledge extraction from the sub-community is easier and faster in fields like social network.A **partition of a community graph** is to divide into clusters(communities), such that each similar vertex belongs to one cluster.In [K5], they have used **Greedy Algorithm** to study a large community graph using compression technique without loss of information or knowledge while compressing.The technique reduces

the iteration steps which may lead for better efficiency which will also help for visualization and analyzing large community graphs.Greedy algorithm is to iteratively group two nodes with the highest cost reduction. It has three phases namely Initialization, Iterative merging and Output.In [K6], they have used Saturated Neighborhood Graph Clustering(SNGC) algorithm which generates efficient clustering to address the issue of unconstrained during construction of the saturated neighborhood graph.It also presents a novel approach to boundary detection in the clustering field.In **SNGC(Saturated Neighbourhood Graph Clustering)** algorithm construction Edge information is constrained on the expansion of key boundary points during the construction of the saturated neighborhood graph.In [G7], they they have used combination of graph similarity metrics and **spectral clustering** algorithm which enables the clients to find similar business models for automated decision making which would help in industry classification for financial reporting.Spectral clustering algorithm is a natural fit for clustering problem, where the data points are clustered directly in the graph space using a graph similarity matrix.In [K8], they have aimed to create an accurate predictive recommendation system for multiplayer RTS games, comparing frequent and discriminative subgraph mining techniques in error rates.**Frequent subgraph mining** technique can be useful to know which subgraphs occur at least n times where n is a user-specified threshold for frequency.In [K9], they have introduced mining techniques for **dynamic graphs** which consists of discovering frequent subgraph sequences,recurrent and triggering patterns and trend sequences.Dynamic graphs are that change over time in terms of attributes (labels) or structures (edges, vertices).In [K10], they have used **possibilistic-clustering** algorithm which helps to identify the cluster(non-terrorist,terrorist-sympathizer,terrorist) that a social network profile has been assigned which helps to find the hints in posts promoting the militants' cause.The possibilistic clustering algorithm, like Fuzzy C-Means, assigns data points degrees of membership to multiple clusters, accommodating uncertainty and allowing for more flexible clustering where points can belong to different clusters with varying degrees of certainty.

It is found from the literature that the applications of Graph mining in the world have become increasingly popular with the rise in the demand for effective data analysis.From research paper [K1-K10] we have explored various diverse aspects of this field.[K1] introduces Random Walk approach for client-oriented data needs,while [K2] enhances accuracy through domain-client relationships,[K3] briefs path for weighted subgraph detection,[K4] partitions community graphs for efficient knowledge extraction and [K5] compresses large graphs using Greedy Algorithm.[K6] addresses unconstrained clustering issue with SNGC,[K7] combines measures and clustering business models,[K8] focuses on RTS game recommendation system comparing error rates,[K9] mines dynamic graphs which has overcame the difficulties of analyzing dynamic graphs in the past and [K10] identifies social network profiles with posts promoting militant causes using possibilistic-clustering.

DATA COLLECTION

The dataset used in this scenario is from SNAP website.To find influential nodes in the Facebook dataset, the study concentrated on a set of edges that came from Facebook profiles that had been anonymised. Participants in the poll provided the data first, which was collected using a special Facebook app. It includes fundamental components like circles, other networks, and node attributes, which stand in for user profiles. New values were used in place of the Facebook internal IDs in order to preserve privacy and safeguard user identities. Furthermore, the dataset's feature vectors were purposefully hidden, making it difficult to identify particular characteristics like political affiliations. The dataset made it possible to analyze user relationships and connections, providing important new information about the dynamics of influence inside the social network.

DATA PREPROCESSING

Data preprocessing is an important first step. It involves loading data from a variety of sources, such as edge and signature data, and organizing it into a format suitable for analysis. Data were transformed to ensure consistency in

the properties of the nodes and to enable accurate comparisons. Additionally, the articles were created by the university. Incomplete or undistributed transaction files to create transaction files. Data preprocessing ensures that the data is consistent and ready for machine learning's next processing.

EXPERIMENTAL RESULTS

Logistic Regression is a powerful machine learning method used for classification tasks. It's applied here to predict whether a node in the network is influential or not based on certain features and labels derived from the network's structure.PageRank is a graph algorithm developed by Google to assess the importance of web pages in their search engine. In this context, it's used to calculate the relevance or importance of each node in the network. The PageRank algorithm considers not only the number of connections a node has but also the quality and importance of those connections.

In this research, we identified the top 10 influential nodes in social data using logistic regression classifiers.The accuracy of this model has a score of 84 and 7.6 this indicates the ability to discriminate between strong networks.We also use PageRank to calculate each site's relevance score. The receiver operating characteristic (ROC) curve shown below shows that the model has good performance in identifying active nodes, with an AUC of approximately [insert AUC value]. The top 10 influential nodes, determined by decision value, are important in the output and provide important information about the actors of the network.

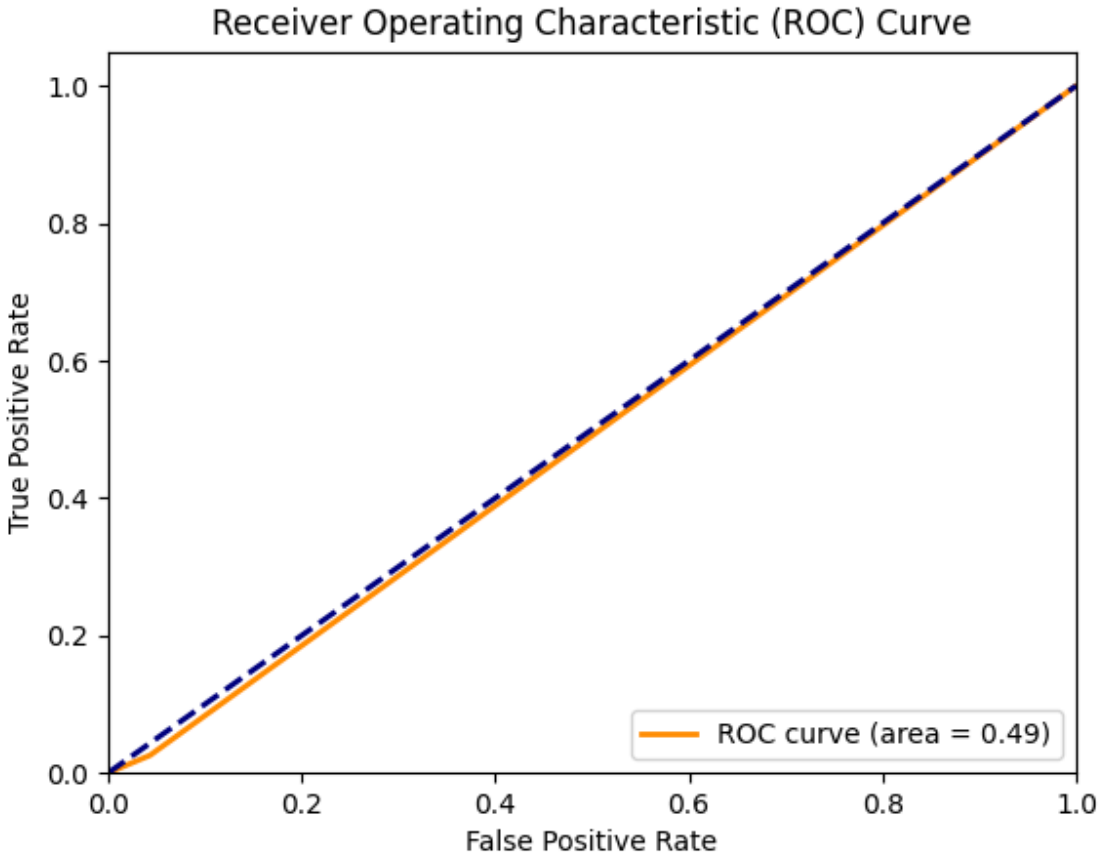

Figure 1:ROC Curve for Pagerank Algorithm

Logistic Regression is a powerful machine learning method used for classification tasks. It's applied here to predict whether a node in the network is influential or not based on certain features and labels derived from the network's structure.Random walk is a versatile algorithm utilized for traversing and exploring networks, providing insights into node importance and connectivity patterns. In this analysis, the code employs random walk scores to assess the significance of each node within the network.

In this study, we found the 10 most common nodes in the network using the random walk scores capability along with logistic regression classifiers. Our model achieved an accuracy of 80%, demonstrating the ability to distinguish important objects in the network. Leveraging the random walk score, this approach increases our access pressure on the network connectivity by providing a better insight into the connector and mediator role of nodes. The receiver operating characteristic (ROC) curve shows the model's ability to detect active nodes, and the nice AUC results indicate that the model is performing well. These top 10 most valuable factors, determined by decision value, give a good idea of the network's key people and their contributions.

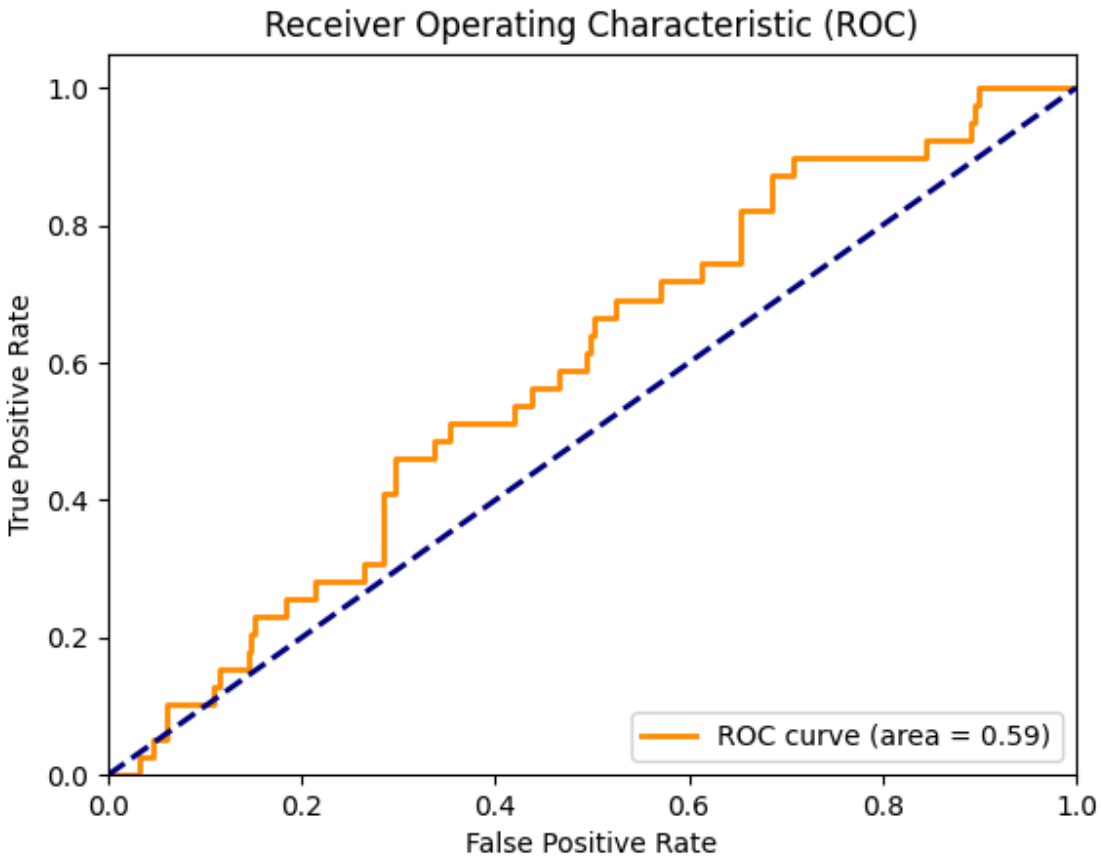

Figure 2:ROC Curve for Random Walk Algorithm

Gaussian Naive Bayes (GNB) is a well-known and widely used classification technique with a unique approach in determining useful parameters in network analysis. GNB's effectiveness in task allocation makes it an important tool for distinguishing dynamic ones in complex networks. In the context of network analysis, GNB evaluates the likelihood of nodes being affected based on specific features and labels derived from the network structure. It assumes a normal distribution of features, hence the name "Gaussian". This assumption simplifies the calculation of the result and reveals the "pure" aspect of the algorithm.

In this study, we use Gaussian Naive Bayes (GNB) to identify the top 10 points in the network. The GNB model showed an admirable accuracy of 52%, demonstrating its ability to distinguish significant values in network patterns. Using the GNB algorithm, this method effectively identifies nodes that act as key connectors and intermediaries, thus revealing their key roles in communication.

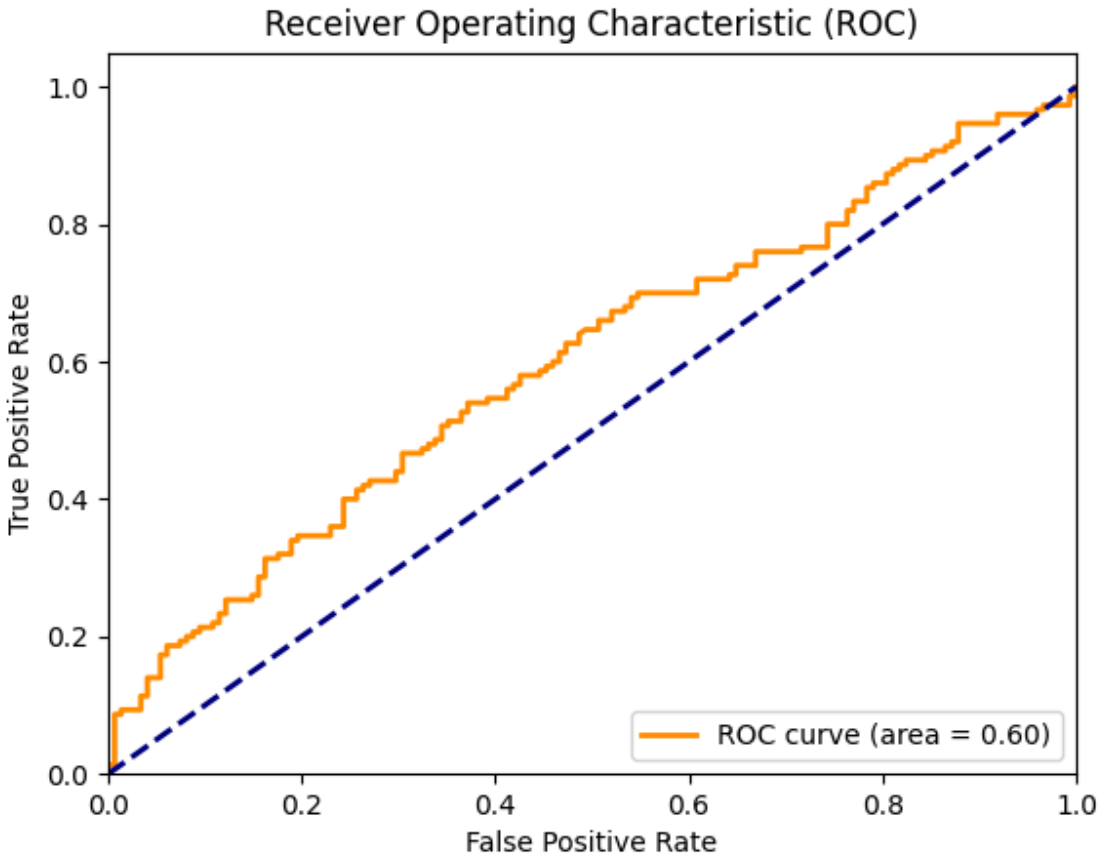

Figure 3:ROC Curve for Gaussian Naive Bayes Algorithm

XGBoost (Extreme Gradient Boosting) is a versatile and powerful learning technique that plays an important role in network analysis and offers a different way to identify influencers. XGBoost is known for its efficiency and effectiveness in complex network operations, making it useful in distinguishing elements in complex networks. In the context of network analysis, XGBoost improves the decision tree and increases the accuracy of prediction by evaluating the importance of nodes based on the unique characteristics and patterns of the network.

In this study, we used the XGBoost function to identify the 10 most relevant networks. The XGBoost model demonstrated its ability to identify significant network patterns, achieving 51% accuracy. By running the XGBoost algorithm, the method effectively identifies nodes that act as important links and intermediaries, demonstrating their important role in supporting communication and information flow.

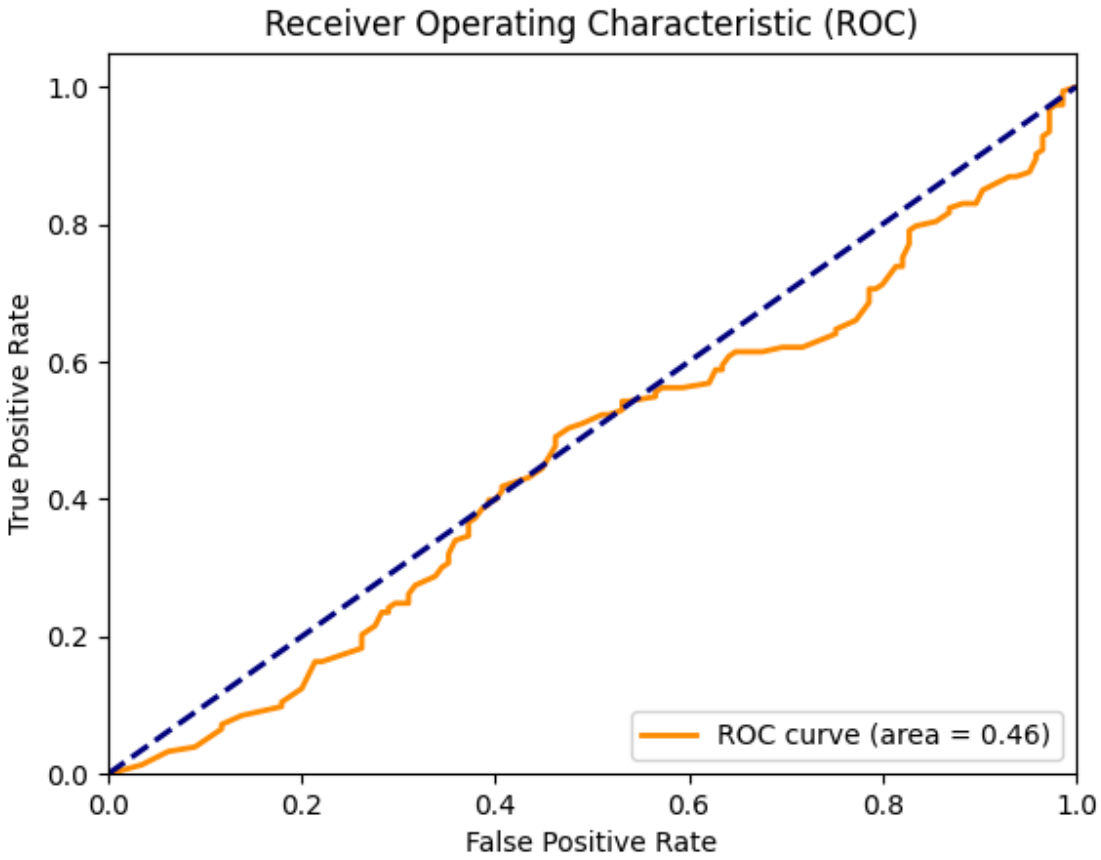

Figure 4:ROC Curve for XGBoost (Extreme Gradient Boosting) Algorithm

Support vector machine (SVM) is a powerful and versatile learning technique frequently used in network analysis. SVM provides a good way to identify influencers or important nodes in complex networks by evaluating decision regions that better distinguish between nodes based on unique features and patterns in the network.

In this study, we use vector machines that support the identification of the 10 most common nodes in the network. The SVM model demonstrated its ability to detect significant connectivity patterns, achieving a classification accuracy of 77%. Using the SVM algorithm, our method effectively identifies nodes that act as important links and intermediaries, highlighting their important role in promoting communication and information flow in the network.

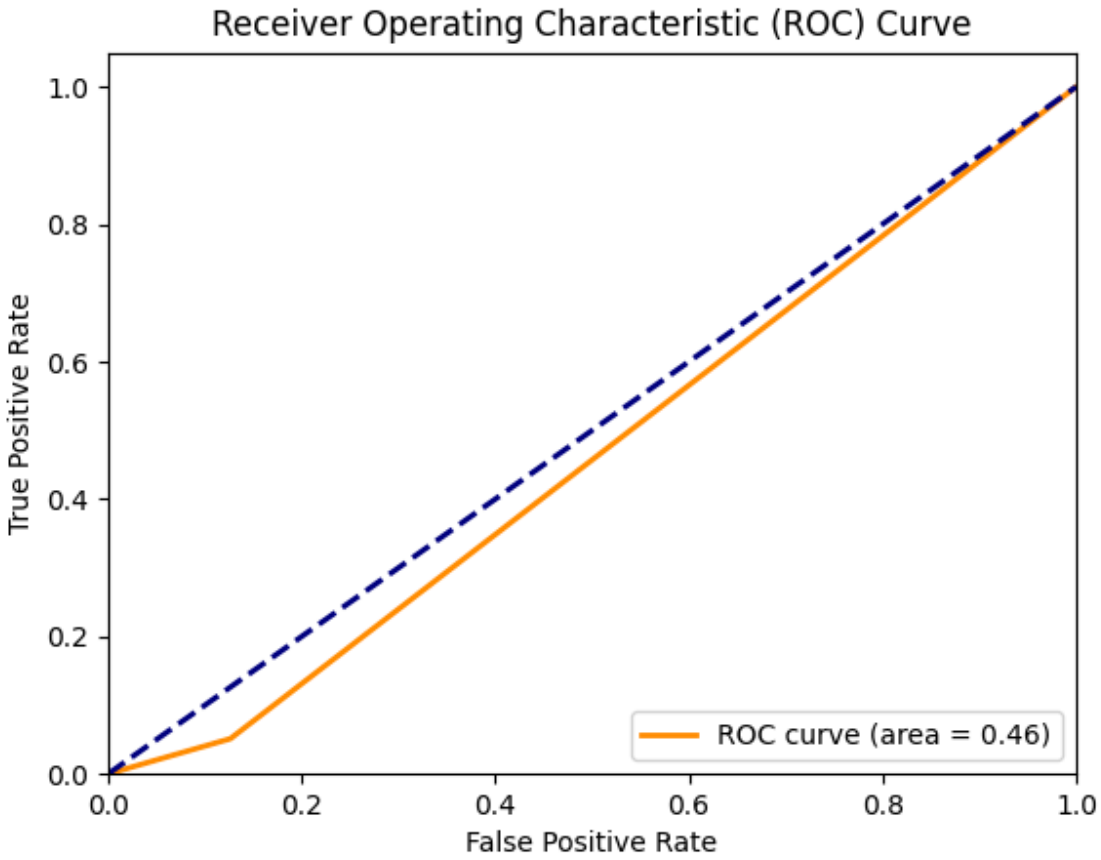

Figure 5:ROC Curve for Support vector machine (SVM) Algorithm

ALGORITHM	ACCURACY	PRECISION	RECALL	MCC	ROC
PageRank	84	76	2.5	2	49.1
Random walk	81	12	5.0	0	59.0
Gaussian Naive Bayes	52	51	92.6	8	60.0
XGBoost	51	52	50.9	2	46.1
Support Vector Machine (SVM)	77	5	5.12	0	46.2

Table1:Performance Metrics Summary

In particular, we are experimenting with five methods that were previously utilized in the research paper that we used as a reference in order to get the best accuracy and precision for a given dataset. We optimized the performance of the five algorithms—Pagerank, Random Walk, Gaussian Naive Bayes, XGBoost, and Support Vector

Machine—by adjusting their hyperparameters. The dataset that is offered consists of a collection of edges derived from anonymized Facebook accounts.This is a general method for determining the optimal algorithm.

1.Data Preprocessing: Do necessary steps for handling missing values, coding of categorical variables, and scaling features.

2.Data splitting: Split the dataset into training and testing sets (e.g. 50% training and 50% testing)

3.Selection models: Train and evaluate various algorithms on datasets. In this example you covered Pagerank, Random Walk, Gaussian Naive Bayes, XGBoost, and Support Vector Machine.Implement each algorithm and evaluate its performance using appropriate metrics.

4.Performance Measurements: When evaluating the performance of an algorithm, choose appropriate metrics that match the performance of the algorithm. Simple metrics for classification include accuracy, which measures overall accuracy; accuracy, which measures the accuracy of forecast quality; return that represents good value; truth and repetition together; ROC AUC, which measures the distribution quality of the model; and the Matthew Strassian Correlation Coefficient (MCC), which measures the performance of classification as a variable. These metrics help determine how well the algorithm performs on a particular task.

5.Hyperparameter Tuning:To maximize the performance of any algorithm, adjust its hyperparameters. The optimal hyperparameters can be found with the aid of methods like random search.

6.Evaluate Models: Utilizing the previously stated metrics, compare each model's performance. Taking into account your unique objectives and needs, select the model that offers the maximum accuracy and precision.

It's crucial to remember that there is no one-size-fits-all solution and that the optimal technique for a particular dataset can change based on the features of the data. To find the best algorithm for a given task, trial and error and thorough analysis are essential. To optimize the performance of each method, adequately preprocess the data and take hyperparameter tweaking into consideration.Through our testing and observations, we were able to conclude that the decision tree method offers higher accuracy and precision than the other algorithms we developed.

COMPARISON

The research paper [K1] takes a similar strategy to ours in identifying important nodes in order to suggest healthy meals based on specific health conditions.Recipes that contain ingredients with "healthy" nutrition nodes may have a higher PageRank score, indicating that they are more suitable for individuals with specific health conditions.Pagerank was calculated for every node according to its centrality in a similar way.Based on node centrality, a variety of methods were explored in both scenarios, with the PageRank approach yielding the best results.

PageRank is a link analysis algorithm used to determine the importance or influence of nodes (e.g., web pages, nodes in a network) in a directed graph or network.

PageRank Formula:

For a node (web page) "A," its PageRank (PR) is calculated as follows:

$$PR(A) = \frac{(1-d)}{N} + \frac{d * \Sigma\,(PR(Ti)}{L(Ti)} \quad \text{……..1}$$

Where:

PR(A): PageRank of node A.

d: Damping factor, a constant between 0 and 1 (typically around 0.85).It represents the probability that a user will continue clicking on links and not jump to a random page.

N: Total number of nodes (web pages) in the network.

Ti: Nodes (web pages) that link to node A.

PR(Ti): PageRank of node Ti.

L(Ti): The number of outbound links on page Ti.

CONCLUSION

In summary, this work explores various algorithms and methods to identify useful nodes in networks and images. By comparing the performance of PageRank, Random Walk, Gaussian Naive Bayes, XGBoost, and Support Vector Machine, we gain a better understanding of their strengths and weaknesses. Among these methods, Pagerank emerged as the most effective method with 84% accuracy and 76% accuracy. Pagerank's transparency and ability to provide detailed explanations make it an important tool for identifying powerful nodes in complex networks.

The ability to identify meaningful nodes in the context of graph mining and network analysis is important for many applications such as relationship analysis, negotiation, and understanding network connections. The information obtained from this research can be useful for decision making and data analysis in many fields.

REFERENCES

[V1] Raza, Shaina, and Chen Ding. n.d. “Fake news detection based on news content and social contexts: a transformer-based approach.” International Journal of Data Science and Analytics 13.4 (2022): 335-362.

[V2] Guo, Muhao, Qiushi Cui, and Yang Weng. n.d. “Graph Mining for Classifying and Localizing Solar Panels in Distribution Grids.” 2023 Panda Forum on Power and Energy (PandaFPE). IEEE, 2023.

[V3] Yoon, Minji, and et al. n.d. “Autonomous graph mining algorithm search with best speed/accuracy trade-off.” 2020 IEEE International Conference on Data Mining (ICDM). IEEE, 2020.

[V4] Zhao, Cheng, and et al. n.d. “Kaleido: An efficient out-of-core graph mining system on A single machine.” 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020.

[V5] Chaudhary, and Hardeo Kumar Thakur. n.d. “Survey of algorithms based on dynamic graph mining.” 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2018.

[V6] Charadava, and Nivid Limbasiya. n.d. “A Review on Various Community Detection Methods in Massive NetworksUsing Graph Mining.” 2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC). IEEE, 2019.

[V7] Tran, Hau, and et al. n.d. “DNS graph mining for malicious domain detection.” 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.

[V8] Liu, Chang, and et al. n.d. “CSRF Detection Based on Graph Data Mining.” 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2020.

[V9] Le, Thanh-Nam, and et al. n.d. “A comic retrieval system based on multilayer graph representation and graph mining.” Graph-Based Representations in Pattern Recognition: 10th IAPR-TC-15 International Workshop, GbRPR 2015, Beijing, China, May 13-15, 2015. Proceedings 10. Springer International Publishing, 2015.

[V10] Yada, Katsutoshi, and et al. n.d. “Consumer behavior analysis by graph mining technique.” Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference, KES 2004, Wellington, New Zealand, September 20-25, 2004, Proceedings, Part II 8. Springer Berlin Heidelberg, 2004.

[J1] Lam, et al. “A graph mining algorithm for classifying chemical compounds.” *2008 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2008*

[J2]Petermann, and André. “Graph mining for complex data analytics.” *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). IEEE, 2016.*

[J3]Shrivastava, et al. “Graph mining framework for finding and visualizing substructures using graph database.” *2009 International Conference on Advances in Social Network Analysis and Mining. IEEE, 2009.*

[J4]Zhang, and Xin. “A General Fault Prediction Framework based on Relationship Mining and Graph Neural Network.” *2022 Global Reliability and Prognostics and Health Management (PHM-Yantai). IEEE, 2022.*

[J5]Liu, and Yang. “BPGM: A Big Graph Mining Tool.” *Tsinghua Science and Technology 19.1 (2014): 33-38.*

[J6]Rao, et al. “An approach to detect sub-community graphs in n-community graphs using graph mining techniques.” *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2016.*

[J7]Rao, et al. “An approach to mining information from telephone graph using graph mining techniques.” *Graph Theoretic Approaches for Analyzing Large-Scale Social Networks. IGI Global, 2018. 34-52*.

[J8]Tran, and Hau. “DNS graph mining for malicious domain detection.” *2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017*.

[J9]Safaei, et al. “Social graph generation & forecasting using social network mining.” *2009 33rd Annual IEEE International Computer Software and Applications Conference. Vol. 2. IEEE, 2009*.

[J10]Nadi, and Farzad. “A new method for mining maximal frequent itemsets based on graph theory.” *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2014*.

[G1] Zhang, Xin, and et al. n.d. “A General Fault Prediction Framework based on Relationship Mining and Graph Neural Network.” *2022 Global Reliability and Prognostics and Health Management (PHM-Yantai). IEEE, 2022.*

[G2] Georgieva-Trifonova, Tsvetanka, and et al. n.d. “Graph-Based Representation of Customer Reviews for Online Stores.” *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, 2019.*

[G3] Shrivastava, Swapnil, and Supriya N. Pal. n.d. “"Graph mining framework for finding and visualizing substructures using graph database.” *2009 International Conference on Advances in Social Network Analysis and Mining. IEEE, 2009.*

[G4] Abdulla, Hasan Hameed Hasan Ahmed, and Husain Hameed Abdulla. n.d. “Fake News Detection: A Graph Mining Approach.” *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD). IEEE, 2023.*

[G5] Guo, Muhao, Qiushi Cui, and Yang Weng. n.d. “Graph Mining for Classifying and Localizing Solar Panels in Distribution Grids.” *2023 Panda Forum on Power and Energy (PandaFPE). IEEE, 2023.*

[G6] Joshi, Anant, and et al. n.d. “An efficient way to find frequent patterns using graph mining and network analysis techniques on the United States airports network.” *Smart Computing and Informatics: Proceedings of the First International Conference on SCI 2016, Volume 2. Springer Singapore, 2018.*

[G7] Wang, Chen, and et al. n.d. “Constraint-based graph mining in a large database.” *Asia-Pacific Web Conference. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.*

[G8] Garza Villarreal,, Sara Elena, and Ramón F. Brena. n.d. “Topic mining based on graph local clustering.” *Advances in Soft Computing: 10th Mexican International Conference on Artificial Intelligence, MICAI 2011, Puebla, Mexico, November 26-December 4, 2011, Proceedings, Part II 10. Springer Berlin Heidelberg, 2011.*

[G9] Yada, Katsutoshi, and et al. n.d. “Consumer behavior analysis by graph mining technique.” *New Mathematics and Natural Computation 2.01 (2006): 59-68.*

[G10] Koga, Hisashi, and et al. n.d. “New application of graph mining to video analysis.” *Intelligent Data Engineering and Automated Learning–IDEAL 2010: 11th International Conference, Paisley, UK, September 1-3, 2010.*

[K1] Kequan, Li, Jiang Z, Wang H, and Liu X. 2020. “Healthy diet recommendation via Food‐Nutrition‐Recipe Graph mining.” 83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

[K2] Tran, Hau, Chuong Dang, Hieu Nguyen, Phong Vo, and Tu Vu. 2019. “Multi-confirmations and dns graph mining for malicious domain detection.” *Intelligent Computing-Proceedings of the Computing Conference. Cham: Springer International Publishing, 2019.*

[K3] Bapuji Rao and Sarojananda Mishra. 2019. “An approach to detect patterns (Sub-graphs) with edge weight in graphs using graph mining techniques.” *Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM 2017. Springer Singapore, 2019.*

[K4] Rao, Bapuji, and Anirban Mitra. 2016. “An algorithm for partitioning community graphs into sub-community graphs using graph mining techniques.” *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics: ICACNI 2015* 2.

[K5] Rao, Bapuji, Mitra Anirban, and D.P.Acharjya. 2015. “A New Approach of Compression of Large Community Graph Using Graph Mining Techniques.” *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2015*.

[K6] Tang, Qi, et al. 2023. “Saturated Neighborhood Graph Clustering Optimization Algorithm Based on Edge Information.” *2023 3rd International Symposium on Computer Technology and Information Science (ICSTIS)*.

[K7] Yang, Steve Y, and et al. n.d. “A graph mining approach to identify financial reporting patterns: an empirical examination of industry classifications.” *Decision Sciences 50.4 (2019): 847-876*.

[K8] Alobaidi, I. A, Leopold, J. L., Allami, A. A., Eloe, N. W., and Tanksley. n.d. “Predictive analysis of real‐time strategy games: a graph mining approach.” *Data Mining and Knowledge Discovery 11.2 (2021): e1398.*

[K9] Fournier‐Viger, and Philippe. n.d. “A survey of pattern mining in dynamic graphs.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.6 (2020): e1372.*

[K10] Moussaoui, Mohamed, Montaceur Zaghdoud, and Jalel Akaichi. n.d. “A possibilistic framework for the detection of terrorism‐related Twitter communities in social media.” *Concurrency and Computation: Practice and Experience 31.13 (2019): e5077.*