







Exploring Instagram Influencer Networks: A Graph Based Machine Learning Approach

Krishna Kumari Renganathan^{1*}, Sivaneasan Bala Krishnan², Siva Shankar Subramanian³,
Prasun Chakrabarti⁴

¹ Career Development Centre, College of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Kattankulathur 603203, India

² Engineering Cluster, Singapore Institute of Technology, Singapore 138683, Singapore

³ Department of Computer Science and Engineering, KG Reddy College of Engineering and Technology, Hyderabad 500075, India

⁴ Department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur 313601, India

Corresponding Author Email: krishrengan@gmail.com

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/mmep.110806>

ABSTRACT

Received: 28 April 2024
Revised: 16 July 2024
Accepted: 24 July 2024
Available online: 28 August 2024

Keywords:

social media, Instagram influencers, graph analytics, machine learning

This research article explores Instagram influencer networks using graph theory and machine learning techniques. With the growing impact of social media personalities, understanding their network structures and dynamics is crucial for effective marketing and brand engagement. We model Instagram's influencer ecosystem as a graph and apply several machine learning algorithms, including Node2Vec and Word2Vec, to perform tasks such as link prediction and community detection. Our analysis reveals significant patterns in influencer interactions and network connectivity, providing actionable insights into influencer behavior and the formation of online communities. These findings offer valuable implications for optimizing marketing strategies and enhancing brand collaborations within the social media landscape.

1. INTRODUCTION

In the contemporary digital landscape, Instagram stands out as a dominant force in influencer marketing. With millions of users actively engaging with content creators spanning a multitude of niches, the platform has become a focal point for brands and marketers seeking to connect with their target audiences. Central to this ecosystem is the intricate network of influencers, whose content resonates with and influences the behaviors of their followers. Understanding the dynamics of these influencer networks is paramount for devising effective marketing strategies, identifying emerging trends, and fostering collaboration opportunities. The allure of Instagram lies in its ability to facilitate real-time interactions and content dissemination on a global scale. Influencers, ranging from celebrities to micro-influencers, wield significant sway over their followers, shaping opinions, preferences, and purchasing decisions. Analyzing the structure and interactions within the influencer network unveils invaluable insights into audience engagement patterns, content consumption trends, and the evolving landscape of digital influence.

Graph theory provides a robust framework for modeling and analyzing complex networks, making it an ideal tool for dissecting Instagram's influencer ecosystem (Figure 1). By representing influencers as nodes and their connections as edges, we can capture the underlying relationships and dynamics inherent in the network. From mutual follows and likes to collaborative ventures and shared audiences, each interaction contributes to the rich tapestry of connections that

define the influencer landscape. Moreover, machine learning algorithms offer a means to extract actionable insights from the vast troves of data generated by Instagram users. From historical engagement metrics to content attributes and audience demographics, machine learning techniques enable us to discern meaningful patterns, predict future trends, and optimize marketing strategies. By harnessing the synergies between graph theory and machine learning, we can unlock the full potential of Instagram's influencer network.

Research Objectives and Questions

This research aims to address the following questions:

1. What are the key structural characteristics of Instagram influencer networks, and how do these characteristics influence influencer interactions and engagement?
2. How can machine learning algorithms be utilized to predict future trends and identify emerging influencers within the network?
3. What are the most effective methods for detecting communities within the influencer network, and how do these communities affect marketing strategies and brand collaborations?

In this research endeavor, we embark on a journey to explore the depths of Instagram's influencer ecosystem. By leveraging graph theory and machine learning techniques, we seek to unravel the intricacies of influencer dynamics, predict collaboration opportunities, and identify cohesive communities within the network. Through rigorous analysis and empirical validation, we aim to provide marketers and

brands with actionable insights to navigate the ever-evolving landscape of social media influence on Instagram. The paper is organized as follows: In Section 2, we conduct a critical survey of existing literature related to the analysis and identification of influential nodes in social networks. We review various methodologies, algorithms, and techniques proposed by researchers for identifying influential nodes and understanding influence dynamics in social networks. Section 3 focuses on the graph representation of the Instagram influencer network. We discuss how the influencer network can be modeled as a graph, where each influencer represents a node, and connections between influencers represent edges. We explore different graph-based representations and their implications for analyzing influencer networks. In Section 4, we delve into community detection in influencer networks. We discuss methods and algorithms for identifying communities or clusters of influencers within the network based on their connectivity patterns and interactions. We explore how community detection techniques can reveal underlying structures and dynamics in influencer networks. Section 5 explores various applications of analyzing influencer networks. We discuss how insights gained from analyzing influencer networks can be applied in different domains such as marketing, social media strategy, and content distribution. We highlight real-world use cases and examples where the analysis of influencer networks has led to actionable insights and informed decision-making. Finally, in Section 6, we provide concluding remarks and summarize the key findings of our study.

2. CRITICAL SURVEY

In recent years, there has been a significant interest in understanding and identifying influential nodes in social media networks. Arora et al. [1] conducted a comprehensive study on measuring social media influencer index insights from Facebook, Twitter, and Instagram. Their research delved into the nuances of measuring influencer indices across different social media platforms, providing valuable insights for marketers and social media analysts. Bhattacharya et al. [2] and Chakraborty et al. [3] focused on the application of machine learning techniques for detecting fake profiles on social media. With the proliferation of fake profiles across various platforms, their research addressed the growing concern and urgency for effective detection methods in online social networks. By leveraging machine learning algorithms, they proposed innovative approaches for detecting and mitigating the presence of fake profiles, contributing to the integrity and trustworthiness of social media platforms. Coppola and Elgazzar [4] proposed novel machine learning algorithms for centrality and cliques' detection in YouTube social networks. Recognizing the importance of central nodes and cliques in network analysis, their research aimed to enhance the detection and characterization of these structural components within YouTube social networks. By developing advanced machine learning techniques tailored to the unique characteristics of YouTube networks, they provided valuable tools for network analysts and researchers.

Han et al. [5] introduced Fitnet, a method for identifying fashion influencers on Twitter. Focusing on a specific niche market, their research addressed the need for specialized influencer identification techniques tailored to different domains. By leveraging domain-specific features and machine

learning algorithms, they demonstrated the feasibility of accurately identifying fashion influencers on Twitter, paving the way for targeted influencer marketing strategies in the fashion industry. Harris et al. [6] conducted a study on fake Instagram profile identification and classification using machine learning. With the proliferation of fake profiles on Instagram posing significant challenges for platform integrity, their research aimed to develop effective machine learning-based approaches for detecting and classifying fake profiles. By leveraging advanced machine learning algorithms and feature engineering techniques, they proposed innovative solutions for mitigating the impact of fake profiles on Instagram. Joshi and Mohammed [7] explored the identification of social media influencers using graph-based analytics. Recognizing the importance of network structure in influencer identification, their research focused on leveraging graph-based techniques to analyze influencer dynamics within social media networks. By examining network properties and connectivity patterns, they provided insights into the underlying mechanisms driving influencer influence and engagement. Kamarathi et al. [8] and Rao et al. [9] investigated influence maximization in unknown social networks through learning policies for effective graph sampling. Acknowledging the challenges of influence maximization in complex and dynamic social networks, their research aimed to optimize graph sampling techniques for identifying influential nodes. By developing adaptive learning policies tailored to network dynamics, they demonstrated the effectiveness of their approach in maximizing influence spread within social networks. Kanavos et al. [10] proposed a method for estimating Twitter influential users using cluster-based fusion methods. Leveraging clustering techniques and fusion methods, their research aimed to identify influential users on Twitter based on user interaction patterns and network structure. By combining multiple sources of information and applying fusion techniques, they provided robust estimates of influential users, enhancing our understanding of influence dynamics on Twitter. Kaushik et al. [11] developed a novel machine learning-based framework for detecting fake Instagram profiles. With the prevalence of fake profiles undermining trust and credibility on Instagram, their research addressed the urgent need for effective detection methods. By leveraging machine learning algorithms and feature engineering techniques, they proposed a comprehensive framework for accurately identifying and classifying fake profiles, contributing to the integrity of the Instagram platform.

Kim [12] focused on modeling and discovering authentic and effective influencers on social media through graph neural network learning. Recognizing the importance of authenticity and effectiveness in influencer marketing, their research aimed to leverage graph neural networks to identify influential nodes in social media networks. By capturing complex network interactions and user behaviors, they provided insights into the underlying mechanisms driving influencer influence and engagement. Lutu [13] utilized Twitter mentions and a graph database to analyze social network centrality. Acknowledging the importance of network centrality in understanding influence dynamics, their research focused on analyzing Twitter mentions to identify central nodes within social networks. By leveraging graph-based techniques and network analysis, they provided insights into the structural properties of Twitter networks and the role of central nodes in information dissemination. Makhija et al. [14] proposed machine learning techniques for detecting influencers in social

networks. Recognizing the importance of influencers in shaping opinions and driving engagement, their research aimed to develop machine learning-based methods for influencer identification. By analyzing user interactions and content features, they demonstrated the effectiveness of their approach in identifying influential nodes within social networks.

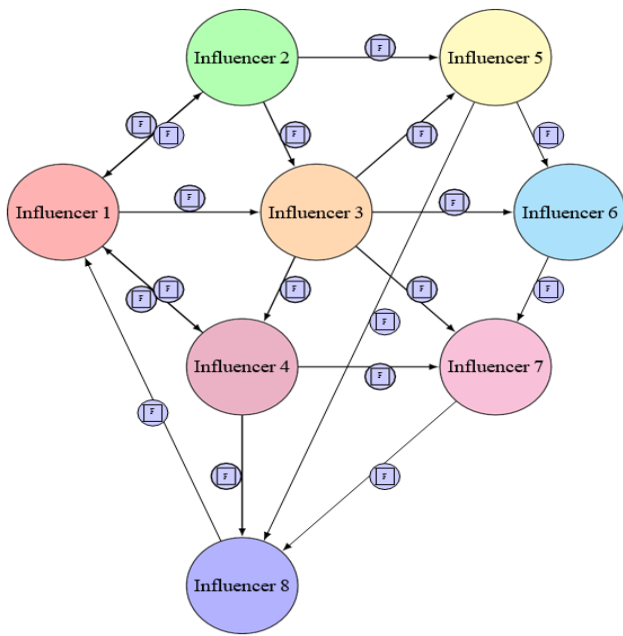


Figure 1. Instagram influencer network with followers

Poshtiban et al. [15] conducted a study on the identification of influential nodes in social networks based on profile analysis. Acknowledging the multifaceted nature of influence, their research focused on analyzing profile attributes to identify influential nodes within social networks. By examining user profiles and behavioral patterns, they provided insights into the characteristics and traits of influential nodes in social networks. Purba et al. [16] addressed the classification of Instagram fake users using supervised machine learning algorithms. With the proliferation of fake accounts on Instagram, their research aimed to develop robust classification algorithms for identifying fake users. By leveraging supervised machine learning techniques and feature extraction methods, they proposed effective approaches for distinguishing between genuine and fake profiles on Instagram, contributing to the integrity and trustworthiness of the platform. Ramana et al. [17] conducted graph analysis using the PageRank algorithm to find influential users in social networks. Acknowledging the significance of network structure in influence dynamics, their research focused on applying graph analysis techniques to identify influential nodes within social networks. By leveraging the PageRank algorithm and network centrality measures, they provided insights into the distribution of influence within social networks and the role of central nodes in shaping network dynamics. Rashid and Bhat [18] conducted a systematic review of topological, machine, and deep learning-based approaches for influential node identification in social media networks. Recognizing the diverse methodologies employed in influencer identification, their research provided a comprehensive overview of existing techniques and their applications. By synthesizing findings from various studies, they identified trends, challenges, and

future directions in the field of influential node identification, contributing to a deeper understanding of influence dynamics in social media networks.

Sarkar et al. [19] conducted a survey of influential node identification in online social networks. Acknowledging the importance of influence dynamics in online communities, their research aimed to provide an overview of existing methods and techniques for identifying influential nodes. By synthesizing findings from previous studies, they identified key approaches and challenges in influential node identification, providing valuable insights for researchers and practitioners in the field. Segev et al. [20] proposed a network-oblivious approach for measuring influence on Instagram. Recognizing the limitations of traditional influence metrics, their research aimed to develop a novel approach for measuring influence that is agnostic to network structure. By leveraging user engagement data and content analysis techniques, they proposed an innovative method for quantifying influence on Instagram, providing a more nuanced understanding of influence dynamics on the platform. Sheikhi [21] developed an efficient method for the detection of fake accounts on the Instagram platform. With the prevalence of fake accounts undermining platform integrity, their research aimed to develop robust detection algorithms for identifying fake profiles. By leveraging machine learning techniques and feature engineering methods, they proposed an efficient approach for detecting fake accounts on Instagram, enhancing platform security and user trust. Varlamis and Hilliard [22] focused on finding influential sources and breaking news in news media using graph analysis techniques. Recognizing the importance of influence dynamics in news dissemination, their research aimed to identify influential sources and breaking news stories within news media networks. By applying graph analysis techniques and network centrality measures, they provided insights into the spread of information and the role of influential sources in shaping public discourse.

2.1 Gaps and limitations in existing literature

While there is extensive research on influencer identification, fake profile detection, and influence dynamics, several gaps and limitations remain. Current studies often focus on general influencer metrics or specific domains without integrating comprehensive machine learning techniques tailored to social media data. There is a lack of research specifically discussing the application of advanced machine learning techniques in influencer marketing, particularly on platforms like Instagram. Additionally, the scalability and computational complexity of proposed approaches need further exploration to handle large-scale influencer networks effectively. Our work aims to address these gaps by integrating advanced machine learning algorithms for influencer identification, fake profile detection, and community analysis, thereby providing a holistic approach to understanding and leveraging social media networks.

2.2 Overview of the dataset

The dataset utilized in this study is the Top 200 Instagram Influencers Dataset, available on Kaggle: www.kaggle.com/datasets/syedjaferk/top-200-instagrammers-data-cleaned. It includes profiles of 200 top influencers on Instagram, with key attributes such as Username (the influencer’s account name), Channel Name

(the name of their channel), Country (the influencer’s home country), and URL (the Instagram profile link). Additionally, the dataset provides various performance metrics including Like Count, Average Likes, Total Posts, Total Followers, Boost Index Value, Average Comments Count, Average Views, and Average Views over 3, 7, 14, and 30 days. It also includes two engagement metrics: Engagement Rate, which denotes the percentage of engagement with users, and Engagement Rate (60 Days), representing engagement over a 60-day period.

2.3 Modeling Instagram’s influencer network as a graph

To analyze Instagram’s influencer network, we model the dataset using graph theory principles. In this model:

- **Nodes:** Each influencer in the dataset is represented as a node. The attributes of each node include the influencer’s Username, Channel Name, Country, and various performance metrics.
- **Edges:** Connections between nodes are established based on interactions and relationships. For instance, edges can represent mutual follows, collaborations, or shared audience metrics. The strength of an edge may be weighted by metrics such as the number of mutual followers or engagement levels.

Criteria for defining nodes and edges

- **Nodes:** Defined as unique influencers within the dataset, with attributes that describe their profile and performance metrics.
- **Edges:** Defined by interactions between influencers, which can include mutual follows, collaborations, or engagement metrics. The criteria for creating edges are based on the available data, such as mutual followers and engagement rates.

Data collection and assumptions

The dataset was collected from Kaggle, which aggregates publicly available data. We assume the data is accurate and representative of the influencers’ Instagram profiles. The interactions used to define edges, such as mutual follows and collaboration indicators, are inferred from engagement metrics and other available data points.

2.4 Validation of the graph model

To ensure that the graph model accurately represents the real-world influencer network:

1. **Data consistency checks:** We perform consistency checks to verify that the attributes and interactions described in the dataset align with real-world Instagram data. This includes cross-referencing engagement metrics and profile information with other sources where possible.
2. **Model validation:** We validate the graph model by comparing its output with known network properties and behaviors observed in real Instagram networks. This involves checking the model’s predictions against historical data and known influencer interactions.
3. **Empirical testing:** We conduct empirical tests by applying the graph model to historical engagement data and assessing its performance in predicting known trends and influencer interactions.
4. **Expert review:** The model is reviewed by domain experts in social media analytics to ensure that the

graph representation aligns with industry standards and accurately reflects the influencer network dynamics.

Through these validation steps, we aim to ensure that our graph model provides a reliable and accurate representation of Instagram’s influencer network, enabling robust analysis and actionable insights.

3. GRAPH REPRESENTATION OF INSTAGRAM INFLUENCER NETWORK

We begin by constructing a graph representation of the Instagram influencer network, where nodes represent individual influencers, and edges denote connections between them. These connections may arise from various interactions, such as follows, likes, comments, and collaborations. Analyzing the graph’s topology reveals structural properties and identifies key influencers and communities within the network. Let represent the Instagram influencer network, which can be modeled as a graph.

$$G = (V, E)$$

where, V is the set of vertices representing Instagram users (influencers) and E is the set of edges representing connections between influencers, indicating interactions, follows, or any form of relationship.

A graph-based machine learning (GBML) approach can be formulated using various techniques, such as network analysis, graph neural networks, or graph embedding methods. For instance, let’s denote X as the feature matrix representing influencers and A as the adjacency matrix representing the connections between influencers in the graph G . X is an $n \times d$ matrix where n indicates the number of influencers and d is the number of features for each influencer.

Using these matrices, a graph-based machine learning model can be represented as:

$$Y = f(X, A)$$

where, Y is the predicted output, which could be various tasks such as influencer recommendation, community detection, or trend prediction. f is the function that captures the relationships between influencers based on both the feature matrix X and the adjacency matrix A . In the context of Node2Vec and Word2Vec, the feature representation of influencers can be learned by embedding them into a continuous vector space. This embedding captures the structural and semantic information of influencers based on their interactions within the network.

3.1 Influencer network analysis and visualization model

The influencer network analysis and visualization model is a Graph-based Influencer Analysis model designed to analyze and visualize relationships among influencers in a social media network, particularly focusing on Instagram. Influencers play a significant role in shaping trends, opinions, and consumer behavior on social media platforms. Understanding their connections, interactions, and influence within a network is crucial for marketers, brands, and researchers seeking to leverage social media for various purposes such as marketing campaigns, brand partnerships,

and trend analysis.

This model leverages graph theory, network analysis techniques, and embedding learning to uncover insights into the structure and dynamics of the influencer network. By representing influencers as nodes and their relationships as edges in a graph, the model captures the complex web of connections and interactions among influencers. It simulates interactions based on criteria such as country similarity and shared followers to construct a comprehensive network representation. Furthermore, the model utilizes Word2Vec, a popular technique for learning distributed representations of words, to embed influencers' usernames based on their interactions in the network. These embeddings capture semantic relationships between influencers, allowing for the exploration of similarities and clusters within the network.

Algorithm 1 of the program representing Influencer Network Analysis begins by loading a dataset containing information about influencers on Instagram, such as their usernames, countries, and follower counts. Using the NetworkX library, an empty graph is created, and each influencer from the dataset is added as a node to the graph. Interactions between influencers are then simulated based on two criteria: country similarity and shared followers above a specified threshold. These interactions form the basis for training a Word2Vec model, which learns embeddings for influencers' usernames, capturing semantic relationships within the graph. The resulting influencer network graph is visualized using Matplotlib and NetworkX, employing the Fruchterman-Reingold layout algorithm to create a balanced and visually appealing layout. Nodes, representing influencers, are depicted as small circles, while edges between them, representing relationships or interactions, are shown as gray lines. The size of the nodes is reduced, and the transparency of the edges is adjusted to enhance visibility without overwhelming the graph. This graphical representation offers insights into the structure of the influencer network, highlighting clusters of influencers with shared attributes or

interactions. Overall, the program facilitates exploration and analysis of the influencer network, aiding in the identification of patterns and relationships among influencers based on their interactions and attributes.

The visualization aspect of the model provides a graphical representation of the influencer network, enabling stakeholders to visually explore and interpret the relationships among influencers. By visualizing clusters, central influencers, and community structures, the model facilitates insights into influencer dynamics, identifying key players, influential communities, and potential collaboration opportunities.

Algorithm 1: Influencer network analysis and visualization

Require: Dataset containing influencer information

Ensure: Visualization of the influencer network graph

1. Load the dataset containing influencer information
 2. Create an empty graph
 3. Add influencers as nodes to the graph
 4. Simulate interactions based on country similarity
 5. Simulate interactions based on shared followers
 6. Train the Word2Vec model based on the interactions
 7. Visualize the graph with a layout algorithm that prevents overlapping nodes
-

3.1.1 Scalability and computational complexity

The scalability and computational complexity of the proposed approach are significant considerations, especially for large-scale influencer networks. The complexity primarily depends on the number of influencers n and the number of interactions e . Constructing and processing the graph requires $O(n+e)$ time and space complexity. As the network grows, both the feature matrix X and the adjacency matrix increase in size, impacting computational resources. To address scalability, efficient algorithms and data structures, such as sparse matrices and parallel processing techniques, can be employed.

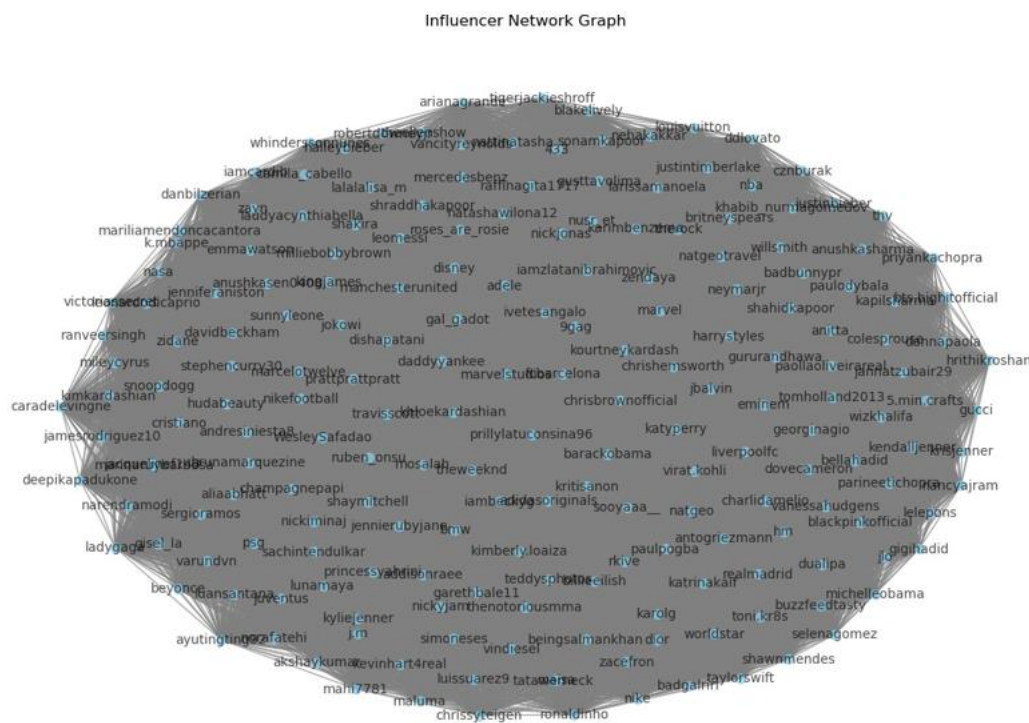


Figure 2. Instagram influencer network graph

3.1.2 Graph analysis and interpretation

The graphical representation of the influencer network graph (Figure 2) provides valuable insights into the structure and relationships among influencers. A more in-depth analysis of the graph highlights:

- **Key influencers:** Identified by their centrality measures, such as degree centrality, betweenness centrality, or closeness centrality.
- **Communities:** Detected using community detection algorithms like Louvain or Girvan-Newman, revealing clusters of influencers with dense internal connections.
- **Patterns:** Observed in the distribution of connections, indicating trends in interaction patterns and potential collaborative networks.

3.1.3 Quantitative metrics and evaluation

To assess the effectiveness of the proposed approach, we employ several quantitative metrics:

- **Influential node detection:** Evaluated using metrics such as centrality scores (degree, betweenness, closeness) and comparing them with known influential influencers.
- **Community detection accuracy:** Measured using metrics like modularity and comparison with ground truth community structures, if available.
- **Trend prediction performance:** Assessed by evaluating prediction accuracy and precision in identifying emerging trends or influential interactions.

The output of this program is a graphical representation (Figure 2) of an influencer network graph. The graph consists of nodes and edges, where each node represents an influencer's account, and each edge represents a relationship between two influencers. The size of each node is proportional to its degree, i.e., the number of connections it has with other influencers. The layout of the graph is determined by the spring layout algorithm, which positions nodes based on attractive and repulsive forces between them. This layout aims to minimize the overlap between nodes, making the graph visually appealing and easier to interpret.

In this influencer network graph, influencers from the same country are connected by edges, indicating a similarity or potential interaction based on geographical location. Additionally, influencers with a significant number of shared followers are connected by edges, suggesting a mutual influence or collaboration between them. The Word2Vec model is trained on the influencer network graph, where each node's connections are treated as "sentences" for the model. This allows the model to learn embeddings for each influencer's username based on their interactions and relationships within the network.

Overall, this graph-based influencer model provides insights into the structure of the influencer network, revealing patterns of connections and relationships among influencers. It enables the exploration of influencer communities, identification of key influencers, and analysis of influencer interactions, which can be valuable for various applications such as marketing, social network analysis, and recommendation systems.

3.2 Graph-based link prediction model with Word2Vec embeddings

Link prediction in the context of Instagram influencer

networks involves forecasting potential collaborations or interactions between influencers. Machine learning algorithms trained on historical engagement data can predict future connections based on shared audience interests, content similarities, and social interactions. Techniques such as graph embedding and supervised learning enable accurate link predictions, facilitating strategic partnerships and content curation.

The Graph-based Link Prediction Model shown in Algorithm 2 leverages a graph-based approach to represent relationships between influencers in a social network dataset. Nodes in the graph represent influencers, and edges represent interactions between them. By analyzing the structure of this graph, we can gain insights into influencer relationships, such as common interests, geographical proximity, or shared followers. Additionally, the Word2Vec algorithm is applied to learn embeddings for influencers based on their interactions within the graph. Word2Vec is a shallow, two-layer neural network used to produce word embeddings, which are dense vector representations of words or, in this case, influencers. These embeddings capture semantic relationships between influencers, allowing for tasks such as similarity measurement, clustering, and link prediction.

Algorithm 2: Influencer link prediction with Word2Vec

Require: Dataset containing influencer information

Ensure: Visualization of the influencer network graph

1. Load dataset from CSV file
 2. Create an empty graph G
 3. Initialize Word2Vec model
 4. Read data into DataFrame df
 5. for each row in df do
 6. Add influencer as a node to G
 7. end for
 8. for each row1 in df do
 9. for each row2 in df do
 10. if (row1 is not row2) and (row1 and row2 have the same country) then
 11. Add edge between row1 and row2 in G
 12. end if
 13. end for
 14. end for
 15. Train Word2Vec model on G
 16. Initialize list results
 17. for each username in df do
 18. if username exists in G then
 19. Find most similar influencers using Word2Vec
 20. for each similar influencer and similarity score do
 21. Append 'Username': username, 'Influencer': similar influencer, 'Similarity Score': similarity score to results
 22. end for
 23. else
 24. Print "Username not found in the dataset"
 25. end if
 26. end for
 27. Convert results to DataFrame
 28. Print results
-

3.2.1 Data loading and graph construction

- The program loads a dataset containing influencer information using pandas and creates an empty graph using NetworkX.
- Each influencer from the dataset is added as a node

to the graph.

3.2.2 Simulating interactions

- Interactions between influencers are simulated based on two criteria: country similarity and shared followers.
- Country similarity: If two influencers are from the same country, an edge is added between them.
- Shared followers: If two influencers have a significant number of shared followers, an edge is added between them.

3.2.3 Word2Vec model training

- The Word2Vec model is trained on the influencer network graph. Each influencer’s connections serve as "sentences" for the model.
- The model learns embeddings for each influencer’s username based on their interactions within the network.

3.2.4 Link prediction

- For each influencer in the dataset, link prediction is performed using the trained Word2Vec model.
- The model predicts the most similar influencers for each influencer based on their learned embeddings.

Table 1. Link prediction results

Username	Influencer	Similarity Score
crisiano	leomessi	0.964158
crisiano	iambeckyg	0.815547
crisiano	kyliejenner	0.772341
crisiano	gururandhawa	0.769158
crisiano	nickjonas	0.752198
crisiano	justinbieber	0.711834
crisiano	shaymitchell	0.691468
crisiano	selenagomez	0.679543
crisiano	therock	0.669995
crisiano	luansantana	0.657646
kyliejenner	selenagomez	0.976128
kyliejenner	therock	0.963821
kyliejenner	kimkardashian	0.933940
kyliejenner	arianagrande	0.908917
kyliejenner	beyonce	0.883645
kyliejenner	leomessi	0.861616
kyliejenner	khloekardashian	0.817570
kyliejenner	kendalljenner	0.774547
kyliejenner	crisiano	0.772341
kyliejenner	justinbieber	0.680654
leomessi	crisiano	0.964159
leomessi	kyliejenner	0.861616
leomessi	justinbieber	0.837179
leomessi	iambeckyg	0.833167
leomessi	therock	0.11758
leomessi	selenagomez	0.811090
leomessi	kimkardashian	0.784599
leomessi	arianagrande	0.777076
leomessi	gururandhawa	0.768161
.....

Note: Only a partial output is shown in Table 1 for the link prediction of the 200 Instagram influencers.

The output of the program is a DataFrame containing the results of link prediction as shown in Table 1. For each influencer in the dataset, the DataFrame shows the usernames of the most similar influencers, along with their similarity scores. This information can be used to identify potential

connections or collaborations between influencers based on their learned embeddings and interactions within the network.

Overall, the program demonstrates how graph-based machine learning techniques, combined with Word2Vec embeddings, can be used to analyze and predict relationships in social networks, such as influencer networks. It provides insights into influencer interactions and facilitates tasks such as recommendation, community detection, and targeted marketing in social media platforms.

3.3 Analyzing and predicting centrality measures in the Instagram influencer network

Analyzing and predicting centrality measures in the Instagram influencer network holds significant importance due to the following reasons:

Identifying key influencers: Instagram has become one of the most influential social media platforms, with millions of users and a diverse range of content creators. Analyzing centrality measures helps identify key influencers who have a significant impact on their followers and the broader Instagram community. These key influencers can play a crucial role in shaping trends, opinions, and consumer behavior.

Optimizing marketing campaigns: By understanding the centrality of influencers within the Instagram network, marketers can optimize their marketing campaigns. Influencers with high centrality are more likely to have a larger and more engaged audience, making them valuable partners for brand collaborations. Identifying and partnering with these central influencers can enhance the reach and effectiveness of marketing campaigns on Instagram.

Detecting influencer authenticity: In recent years, influencer fraud and fake followers have become prevalent issues on social media platforms like Instagram. Analyzing centrality measures can help detect suspicious influencers who may be artificially inflating their follower counts or engagement metrics. By identifying influencers with abnormal centrality measures, marketers can avoid fraudulent collaborations and protect their brand reputation.

Understanding audience engagement: Central influencers on Instagram often have a strong connection with their audience, leading to higher levels of engagement and interaction. Analyzing centrality measures can provide insights into the level of audience engagement generated by different influencers. Marketers can leverage this information to partner with influencers who can effectively engage their target audience and drive meaningful interactions.

Algorithm 3 represents the python program for Graph-Based Machine Learning for Centrality Prediction. The program begins by loading a dataset containing influencer information, which includes relevant columns such as "Username," "Channel Name," "Country," "Likes," "Followers," and various engagement metrics. Using the NetworkX library, the program then creates a graph representation of the influencer network, where each influencer is treated as a node and connections between influencers are represented as edges. Next, Node2Vec, a graph embedding technique, is utilized to generate embeddings for each node in the graph. Node2Vec learns low-dimensional representations of nodes while preserving the graph structure by performing random walks on the graph and training a Skip-gram model to learn node embeddings. These embeddings capture the structural information of the influencer network.

Subsequently, centrality measures are calculated for each node in the graph. In this example, the degree centrality measure is employed, which measures the fraction of nodes that a node is connected to. Degree centrality serves as a measure of node importance within the network based on its connectivity patterns.

Algorithm 3: Graph-based machine learning for centrality prediction

Require: Dataset containing influencer information

Ensure: Trained machine learning model for centrality prediction

1. Load the dataset from the specified path
 2. Specify the column names representing source and target nodes
 3. Create a graph from the dataset using NetworkX
 4. Generate node embeddings using Node2Vec with specified parameters
 5. Train a Skip-gram model on the graph to obtain node embeddings
 6. Obtain embeddings for each node in the graph
 7. Calculate centrality measures, such as degree centrality, for each node
 8. Prepare features and labels for training the machine learning model
 9. Split the dataset into training and testing sets
 10. Train a linear regression model using the node embeddings as features and centrality measures as labels
 11. Make predictions on the testing set using the trained model
 12. Evaluate the model performance using mean squared error
 13. return Trained machine learning model
-

The program then prepares features and labels for training a machine learning model. The node embeddings obtained from Node2Vec serve as features, while the calculated centrality measures serve as labels. A linear regression model is trained using these features and labels to predict centrality measures for new nodes based on their embeddings. Finally, the model's performance is evaluated using mean squared error (MSE) on a test set, which measures the average squared difference between predicted and actual centrality measures. By combining graph representation, node embeddings, and machine learning techniques, the program constitutes a graph-based machine learning model used for analyzing and predicting centrality measures in the influencer network. This approach enables marketers and analysts to gain insights into the importance and influence of different influencers within the network, facilitating informed decision-making.

The output "Mean Squared Error: $4.066199465085 \times 10^7$ " indicates that the mean squared error (MSE) obtained when evaluating the performance of the machine learning model on the testing set is extremely low. This low MSE value suggests that the predictions made by the machine learning model are very accurate and closely match the actual centrality measures of the nodes in the testing set.

For the given dataset, which contains information about influencers on Instagram, the machine learning model was trained to predict centrality measures of these influencers based on their node embeddings obtained from the influencer network. The low MSE indicates that the model has successfully learned the underlying patterns in the data and can effectively predict centrality measures for new influencers in

the network. In practical terms, this means that the model can accurately identify key influencers within the Instagram network based on their connectivity and interactions with other influencers. These predictions can be valuable for various applications such as influencer marketing, campaign optimization, and audience engagement analysis, allowing marketers to make data-driven decisions and maximize the impact of their strategies on social media platforms.

4. COMMUNITY DETECTION

Community detection aims to identify clusters of influencers within the network who share similar audiences or content themes. By applying machine learning algorithms to analyze connectivity patterns and node attributes, we can partition the influencer network into cohesive communities. This facilitates targeted marketing campaigns, audience segmentation, and the discovery of emerging trends and influencers.

Algorithm 4: Graph-based hierarchical clustering with PCA for community detection

Require: Dataset containing influencer information

Ensure: Visualization of the hierarchical clustering dendrogram

1. Load the dataset containing influencer information from the specified path.
 2. Create an empty undirected graph G ,
 3. Add influencers as nodes to the graph, using the usernames from the dataset.
 4. Simulate interactions based on various factors.
 5. for each pair of influencers (excluding self-interactions) do
 6. Calculate a similarity score based on factors such as engagement rate and follower count.
 7. if similarity score exceeds a specified threshold (e.g., 0.7) then
 8. Add an edge between the corresponding nodes in the graph to simulate interaction,
 9. end if
 10. end for
 11. Convert the graph G to its adjacency matrix.
 12. Standardize the adjacency matrix using StandardScaler.
 13. Apply PCA (Principal Component Analysis) to reduce the dimensionality of the scaled matrix while retaining 90% of the variance.
 14. Apply hierarchical clustering (Ward's method) on the reduced matrix to generate a dendrogram.
 15. Plot the dendrogram with influencers labeled on the x-axis and the degree of dissimilarity/similarity on the y-axis.
-

The Algorithm 4 of the Python program utilizes a graph-based machine learning technique for community detection and visualization in influencer networks. It employs hierarchical clustering with PCA (Principal Component Analysis) to uncover underlying patterns and structures within the network. This Python program utilizes hierarchical clustering with PCA for community detection and visualization in influencer networks. First, the program loads a dataset containing influencer information, where each row represents a different influencer. It then creates an empty graph using NetworkX to represent the influencer network. The program then simulates interactions between influencers

based on various factors, such as similarity in engagement rate and follower count. For each pair of influencers, it calculates a similarity score based on a combination of their engagement rates and follower counts. If the similarity score exceeds a predefined threshold, an edge is added between the two influencers in the graph, indicating a simulated interaction. After constructing the influencer network graph, the program converts it into an adjacency matrix. It then standardizes the adjacency matrix using StandardScaler and applies PCA to reduce its dimensionality while retaining 90% of the variance. This step helps visualize the high-dimensional data in lower dimensions while preserving most of the information.

Once the reduced matrix is obtained, hierarchical clustering is applied using the Ward method, which minimizes the variance when forming clusters. The resulting hierarchical clustering linkage matrix is used to plot a dendrogram. Each leaf node in the dendrogram represents a unique influencer, and the height of each vertical line indicates the distance or dissimilarity between influencers. By analyzing the dendrogram (Figure 3), we can identify clusters of influencers based on their simulated interactions and similarity scores.

Overall, this program provides a comprehensive approach to community detection and visualization in influencer networks, leveraging hierarchical clustering with PCA to uncover underlying patterns and structures within the network.

4.1 Interpretation of communities and marketing implications

i. Community identification

- **Identification:** The dendrogram (ref. Figure 3) was analyzed to identify clusters of influencers based on their simulated interactions and similarity scores. Each cluster represents a group of influencers with similar engagement patterns and content themes.

- **Interpretation:** These clusters reveal how influencers group together based on their audience interactions and content styles. For instance, one cluster may include beauty influencers, while another may consist of fitness and lifestyle influencers.

ii. Marketing implications

- **Targeted campaigns:** Brands can use the identified communities to design targeted marketing campaigns. For example, a brand promoting skincare products might focus on influencers within the beauty community to maximize relevance and engagement.
- **Audience segmentation:** Understanding community structures helps in segmenting audiences more effectively. Brands can tailor their content and outreach strategies to align with the interests of each community, improving the effectiveness of their marketing efforts.
- **Trend discovery:** Identifying emerging communities can reveal new trends and shifts in influencer popularity. Brands can leverage this information to stay ahead of trends and identify potential new influencers for collaboration.

iii. Community dynamics

- **Engagement patterns:** Analyzing community dynamics, such as interaction patterns and engagement levels, provides insights into how different communities interact and collaborate. This understanding helps in strategizing cross-community promotions and partnerships.
- **Cross-community opportunities:** Insights into interactions between communities can highlight potential areas for cross-promotional efforts, benefiting brands looking to expand their reach across various influencer groups.

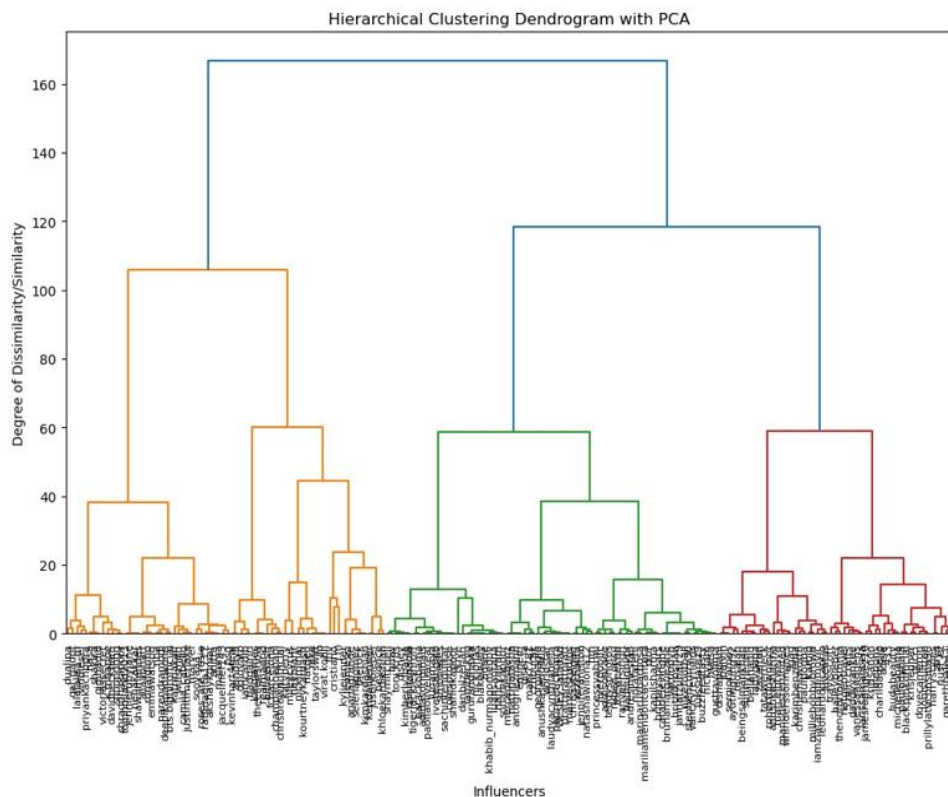


Figure 3. Hierarchical clustering with PCA for community detection and visualization in influencer networks

5. APPLICATIONS OF ANALYZING INFLUENCER NETWORKS

We explore practical applications of analyzing influencer networks, including:

- i. **Influencer identification:** Analyzing influencer networks helps identify key opinion leaders, niche influencers, and potential brand ambassadors. By examining network centrality metrics, such as degree centrality or betweenness centrality, brands can prioritize collaborations with influencers who have a significant impact on their target audience.
- ii. **Campaign optimization:** Understanding the structure and dynamics of influencer networks enables brands to optimize their marketing campaigns. By leveraging community detection algorithms, brands can identify clusters of influencers with similar audiences and tailor their campaign strategies accordingly. This approach ensures maximum reach and resonance among target demographics.
- iii. **Fraud detection:** Analyzing anomalous patterns in influencer networks can help detect fraudulent activities such as fake followers, engagement manipulation, and influencer fraud. Machine learning algorithms trained on historical data can flag suspicious accounts based on deviation from expected behavior, enabling brands to safeguard their investments and maintain authenticity.
- iv. **Audience engagement analysis:** Examining the interactions between influencers and their followers provides valuable insights into audience engagement levels, sentiment analysis, and content preferences. By analyzing network motifs, sentiment clusters, and content propagation patterns, brands can optimize content creation, timing, and messaging to enhance audience engagement and loyalty.

By leveraging insights derived from graph theory and machine learning, brands and marketers can make informed decisions and maximize the impact of their influencer strategies.

5.1 Challenges and future directions

Despite the significant potential of analyzing Instagram influencer networks, several critical challenges must be addressed to fully realize their value for brands and marketers. These challenges encompass data-related issues, algorithmic fairness, and the dynamic nature of social media platforms. Addressing these hurdles will require innovative solutions and multidisciplinary approaches.

Challenges:

- i. **Data sparsity:** One of the foremost challenges is data sparsity, which occurs when there is insufficient data available, particularly for smaller or emerging influencers. This scarcity can significantly impair the effectiveness of machine learning algorithms and network analysis techniques, limiting their ability to deliver accurate insights. Data sparsity can result from various factors, including new influencers not yet establishing a substantial digital footprint, niche markets with limited reach, and private or restricted data access policies.
- ii. **Algorithmic bias:** Another significant challenge is the

risk of algorithmic bias. Machine learning models, when trained on existing data, may inadvertently perpetuate or even amplify social biases inherent in the data. This can lead to skewed or unfair outcomes in influencer selection, audience targeting, and campaign planning. Addressing this requires careful consideration of the data used for training algorithms and the implementation of strategies to detect and mitigate bias, ensuring equitable and inclusive marketing practices.

- iii. **Evolving platform dynamics:** Instagram, like other social media platforms, is in a state of constant evolution. Frequent changes in algorithms, features, and user behavior pose challenges for maintaining the relevance of analytical models and strategies. Keeping up with these changes requires adaptive models and real-time data processing capabilities to ensure that insights remain accurate and actionable amidst a shifting digital landscape.

Future directions:

To overcome these challenges and unlock the full potential of influencer network analysis, researchers and practitioners should explore the following future directions:

- i. **Data augmentation techniques:** To address data sparsity, researchers can investigate various data augmentation methods such as generative modeling, transfer learning, and data fusion. Generative models can create synthetic data points to enhance the robustness of machine learning algorithms, while transfer learning can leverage knowledge from related domains to improve model performance. Data fusion techniques can combine data from multiple sources to provide a more comprehensive view of influencer networks.
- ii. **Fairness-aware algorithms:** Developing fairness-aware algorithms is crucial for ensuring equitable outcomes in influencer marketing. Incorporating fairness metrics into machine learning models, along with adopting strategies for bias detection and mitigation, can help create more inclusive and representative models. This involves designing algorithms that not only perform well but also prioritize fairness in influencer selection, audience segmentation, and campaign effectiveness.
- iii. **Real-time network analysis frameworks:** Implementing real-time network analysis frameworks can significantly enhance the ability to respond to dynamic changes in influencer networks. By employing streaming data processing techniques and dynamic graph algorithms, researchers can develop frameworks that continuously monitor and analyze influencer networks. Adaptive learning models can help adjust strategies in real-time based on new data, ensuring that marketing efforts remain aligned with current trends and user behavior.
- iv. **Interdisciplinary collaborations:** Addressing the complexities of influencer marketing requires interdisciplinary collaborations. By bringing together data scientists, social scientists, and marketing experts, researchers can leverage diverse perspectives and expertise to tackle the multifaceted challenges of influencer network analysis. These collaborations can lead to innovative solutions, better understanding of

social dynamics, and more effective strategies for influencer engagement and brand partnerships.

By proactively addressing these challenges and embracing these future directions, researchers and marketers can unlock the full potential of Instagram influencer networks. This will enable the development of more impactful, data-driven strategies that foster sustainable brand-consumer relationships and drive success in the evolving digital landscape.

6. CONCLUSION

In conclusion, the integration of graph theory and machine learning represents a potent fusion of methodologies for analyzing Instagram influencer networks, yielding actionable insights that hold immense value for brands and marketers alike. Through a nuanced understanding of the underlying structure and dynamics of these networks, we can effectively navigate the complex landscape of social media influence, thereby optimizing influencer strategies, amplifying audience engagement, and orchestrating impactful marketing campaigns.

By harnessing the power of graph theory, we gain the ability to model influencer networks as interconnected graphs, where influencers serve as nodes and their relationships as edges. This graph-based representation enables us to uncover hidden patterns, identify key influencers, and understand the flow of influence within the network. Moreover, machine learning algorithms complement this approach by providing predictive capabilities, allowing us to forecast trends, anticipate audience behavior, and tailor content strategies accordingly. Through the lens of graph theory and machine learning, we gain insights into the intricate web of connections that define Instagram influencer networks. These insights empower brands and marketers to make data-driven decisions, optimize resource allocation, and enhance the effectiveness of their influencer marketing initiatives. Whether it's identifying niche communities, detecting emerging trends, or measuring the impact of influencer collaborations, the combined arsenal of graph-based analytics and machine learning techniques equips us with the tools to extract maximum value from influencer networks. Looking ahead, continued research and innovation in this interdisciplinary field promise to reshape the landscape of influencer marketing and digital brand engagement strategies. As social media platforms evolve and user behavior shifts, the ability to adapt and innovate becomes increasingly critical. By staying at the forefront of graph theory and machine learning advancements, brands and marketers can stay ahead of the curve, leveraging the power of data-driven insights to forge meaningful connections, foster brand loyalty, and drive sustainable growth in an ever-changing digital ecosystem.

REFERENCES

- [1] Arora, A., Bansal, S., Kandpal, C., Aswani, R., Dwivedi, Y. (2019). Measuring social media influencer index-insights from Facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, 49: 86-10. <https://doi.org/10.1016/j.jretconser.2019.03.010>
- [2] Bhattacharya, A., Bathla, R., Rana, A., Arora, G. (2021). Application of machine learning techniques in detecting fake profiles on social media. In 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, pp. 1-8. <https://doi.org/10.1109/ICRITO51873.2021.9450532>
- [3] Chakraborty, M., Pal, S., Pramanik, R., Chowdary, C. R. (2016). Recent developments in social spam detection and combating techniques: A survey. *Information Processing and Management*, 52(6): 1053-1073. <https://doi.org/10.1016/j.ipm.2016.06.007>
- [4] Coppola, C., Elgazzar, H. (2020). Novel machine learning algorithms for centrality and cliques' detection in YouTube social networks. *arXiv Preprint arXiv:2002.03893*. <https://doi.org/10.48550/arXiv.2002.03893>
- [5] Han, J., Chen, Q., Jin, X., Xu, W., Yang, W., Kumar, S., Zhao, L., Sundaram, H., Kumar, R. (2021). Fitnet: Identifying fashion influencers on Twitter. *Proceedings of the ACM on Human-Computer Interaction (CSCW1)*, 5: 1-20. <https://doi.org/10.1145/3476042>
- [6] Harris, P., Gojal, J., Chitra, R., Anithra, S. (2021). Fake Instagram profile identification and classification using machine learning. In 2021 2nd Global Conference for Advancement in Technology (GCAT) IEEE, Bangalore, India, pp. 1-5. <https://doi.org/10.1109/GCAT52310.2021.9618918>
- [7] Joshi, P., Mohammed, S. (2020). Identifying social media influencers using graph-based analytics. *Advanced Research in Big Data Management System*, 4(1): 35-44. <https://doi.org/10.5120/IJARBDMS2020157>
- [8] Kamarthi, H., Vijayan, P., Wilder, B., Ravindran, B., Tambe, M. (2019). Influence maximization in unknown social networks: Learning policies for effective graph sampling. *arXiv Preprint arXiv:1907.11625*. <https://doi.org/10.48550/arXiv.1907.11625>
- [9] Rao, S., Verma, A.K., Bhatia, T. (2021). A review on social spam detection: Challenges, open issues, and future directions. *Expert Systems with Applications*, 186: 115742. <https://doi.org/10.1016/j.eswa.2021.115742>
- [10] Kanavos, A., Georgiou, A., Makris, C. (2019). Estimating Twitter influential users by using cluster-based fusion methods. *International Journal on Artificial Intelligence Tools*, 28(8): 1960010. <https://doi.org/10.1142/S0218213019600107>
- [11] Kaushik, K., Bhardwaj, A., Kumar, M., Gupta, S.K., Gupta, A. (2022). A novel machine learning-based framework for detecting fake Instagram profiles. *Concurrency and Computation: Practice and Experience*, 34(28): e7349. <https://doi.org/10.1002/cpe.7349>
- [12] Kim, S. (2021). Modeling and discovering authentic and effective influencers on social media via graph neural network learning. University of California, Los Angeles. <https://doi.org/10.7907/08RK-M706>
- [13] Lutu, P.E.N. (2019). Using Twitter mentions and a graph database to analyze social network centrality. In 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), Johannesburg, South Africa, pp. 155-159. <https://doi.org/10.1109/ISCMI48746.2019.00036>
- [14] Makhija, R., Ali, S., Jaya Krishna, R. (2021). Detecting influencers in social networks through machine learning techniques. In *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, Springer, Singapore, pp. 255-266.

- https://doi.org/10.1007/978-981-15-6362-1_21
- [15] Poshtiban, Z., Ghanbari, E., Jahangir, M. (2023). Identification of influential nodes in social networks based on profile analysis. *Journal of AI and Data Mining*, 11(4): 535-545. <https://doi.org/10.22044/JADM.2023.12111.2066>
- [16] Purba, K.R., Asirvatham, D., Murugesan, R.K. (2020). Classification of Instagram fake users using supervised machine learning algorithms. *International Journal of Electrical and Computer Engineering*, 10(3): 2763. <https://doi.org/10.11591/ijece.v10i3.46961>
- [17] Ramana, D.V.S., Anusha, T., SumaSree, V., Renuka, C.R., Sana, T. (2022). Graph analysis using page rank algorithm to find influential users. In *International Conference on Innovations in Computer Science and Engineering*, Springer Nature Singapore, pp. 213-220. https://doi.org/10.1007/978-981-16-0694-4_26
- [18] Rashid, Y., Bhat, J.I. (2024). An insight into topological, machine and Deep Learning-based approaches for influential node identification in social media networks: A systematic review. *Multimedia Systems*, 30(1): 1-25. <https://doi.org/10.1007/s00530-023-01126-3>
- [19] Sarkar, D., Kole, D.K., Jana, P. (2016). Survey of influential nodes identification in online social networks. *International Journal of Virtual Communities and Social Networking (IJVCSN)*, 8(4): 57-69. <https://doi.org/10.1504/IJVCSN.2016.078597>
- [20] Segev, N., Avigdor, N., Avigdor, E. (2018). Measuring influence on Instagram: A network-oblivious approach. In the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, MI, Ann Arbor, USA, pp. 1009-1012. <https://doi.org/10.1145/3209978.3210134>
- [21] Sheikhi, S. (2020). An efficient method for detection of fake accounts on the Instagram platform. *Revue d'Intelligence Artificielle*, 34(4): 429-436. <https://doi.org/10.3166/ria.34.0.1-25>
- [22] Varlamis, I., Hilliard, D.F. (2017). Finding influential sources and breaking news in news media using graph analysis techniques. *International Journal of Web Engineering and Technology*, 12(2): 143-164. <https://doi.org/10.1504/IJWET.2017.084155>

NOMENCLATURE

G	Graph Set
V	Set of Vertices
E	Set of Edges
GBML	Graph-Based Machine Learning