Role of Influencers in Paid Promotions: Similarity Calculations and Community Detection

Yash Kangralkar[1], Shradha I. Bavalatti[1], Nikhil Bavkar[1], Shreya Ankale[1], and Santosh Pattar[1]

Department of Computer Science and Engineering,
KLE Technological University's
Dr. M. S. Sheshgiri College of Engineering and Technology
yashkangralkar2510@gmail.com

Abstract. Social Network Analysis (SNA) has emerged as a significant research field, that helps in understanding interpersonal relations and information dissemination within the social network. Community detection plays a pivotal role in SNA identifying cohesive groups of individuals which in turn helps in identifying the underlying structure, analyzing patterns, and facilitating targeted interventions in various domains. In this paper, a novel community detection technique based on similarity matches is presented. We propose a similarity-based score that leverages the inherent structures of the social network to detect communities and carry out impact analysis. We demonstrate the effectiveness of our algorithm by comparing it with Machine Learning (ML) techniques. Furthermore, we discuss the implications of our work for real-world applications highlighting its potential for enhancing community detection in diverse domains, in particular for paid promotion tweet analysis.

Keywords: Social network analysis · Community detection · Directed acyclic graphs · Influencer · Node similarity

1 Introduction

A social network is a web of social connections between various humans. The analysis of these complex domains is done by studying the structure and dynamics of the network through Social Network Analysis (SNA). In SNA, the users or the individual entities are represented as nodes, and the unique interactions or relations by edges. In the modern days social network plays a vital role in an individual's decision-making process, in various commercial domains like advertisement, shopping, *etc* [8]. This surge of social media is dominated mainly by influencers, who are responsible for actively manipulating the network dynamics through targeted content creation and strategic community engagement. With the help of SNA, we can enhance these diffusions of information and explore the power of influencers in shaping the social network domain [4].

Influencers, who are individuals with significant social followings and innovative ideas for content creation, wield enormous influence over their target audience. Therefore, the rise of influencer communities has seen wide growth as they can collectively leverage the network effects to amplify their reach and foster a sense of trust and reliability[3]. In SNA, various metrics are utilized to identify specific communities and their influential individuals. Hence, community detection helps to discover the hidden structures within these networks by recognizing clusters of influencers who share common interests, audiences, and collaborations [5].

In the domain of community detection concerning influential entities, a wide spectrum of methodological approaches has been deployed to deepen our understanding of social network dynamics. In the recent past, several techniques like fuzzy concept analysis (Rios *et. al.,* [9]), network representation learning (Li *et. al.,* [6]), transfer entropy statistical causality method (Chikhaoui *et. al.,* [2]) and personalized page rank algorithms (Alp *et. al.,* [10]) to gain SNA insights. Similarly, these studies represent a fraction of the diverse approaches in the existing works, underscoring the breadth and depth of inquiry into social network dynamics and influence mechanisms.

Our approach for community detection centers around a novel similarity calculation approach. By thorough review of existing literature on community detection and impact analysis, we identified the gaps. To address these gaps we consider various parameters while designing our similarity function. Thus, the contributions of this work are as follows.

- *Similarity Algorithm for Community Detection:* We introduce an innovative similarity algorithm for community detection.
- *Impact Analysis of Influencers:* In-depth analysis of influencers' impact within social networks.
- *Comparison with Machine Learning Techniques:* We compare the proposed community detection technique within the clustering algorithm to determine its effectiveness.

The rest of the paper is organized as follows. Section 2, provides an overview of the existing works in the field of community detection and impact analysis. In the next Section 3, we define our problem statement and methodology for community detection and influencer impact analysis. In Section 4, we present our experimentation and results. Finally, concluding remarks are presented in Section 5.

2 Literature Survey

In this section, we survey recent works on community detection and further interpret their advantages and disadvantages.

Rios *et. al.,* [9] proposed a methodology to screen out irrelevant information using semantic analysis. This methodology is based on fuzzy concept analysis and latent Dirichlet analysis. It is advantageous as it detects influencers with a higher

level of accuracy. However, methods for computing precise semantic distance are not discussed. Li *et. al.,* [6] came up with an effective node representation strategy. It examines individual node's influence on information and community affiliation. With the proposed methodology the quality of the prediction node representation is improved. However, the model is not tested for larger datasets.

Alp *et. al.,* [10] proposed a methodology named personalized page rank. This algorithm incorporates data obtained from the network topology of user actions, that leads to the determination of topical influencers specializing in certain fields. The usage of large data sets enhanced the efficiency of the proposed work. However, all the features were not considered. Arora *et. al.,* [1] came up with a mechanism for computing the influencer index on famous social media platforms Twitter, Facebook, and Instagram. The proposed model uses a regression approach and includes a set of 39 features that help in determining the impact on the consumers. Later, these features are analyzed and a cumulative score is obtained. This method outperforms various approaches based on error rate and accuracy. However, identifying the personality types of the influencer is not discussed.

Logan *et. al.,* [7] came up with a directed multilayer network approach. The proposed work is effective in identifying influencers and communities *via.* broad query enabling high coherence. However, latent Dirichlet analysis confronted difficulties in identifying diverse topics of the conversation. This leads to the degradation of the network layer. Zheng *et. al.* [11] proposed an on-demand influencer discovery framework that identifies influencers. This model utilizes a iterative learning approach that incorporates the language attention network as a subject filter and employs the influence convolution network, which is based on user interactions. Evaluations on Twitter datasets have shown a good ratio of related tweets and it detects topic-specific influencers.

Table 1: Evaluation of Recent Works on Community Detection.

Author	**Approach**	**1**	**2**	**3**	**4**	**5**[1]
Rios *et. al.* [9]	Fuzzy Concept Analysis	×	×	×	✓	×
Li *et. al.* [6]	Network Representation Learning	✓	✓	×	✓	✓
Alp *et. al.* [10]	Personalized Page Rank	✓	✓	×	×	×
Arora *et. al.* [1]	Regression Models	×	✓	✓	✓	✓
Logan *et. al.* [7]	Latent Dirichlet Analysis	✓	✓	✓	×	✓
Zheng *et. al.* [11]	On-Demand Influencer Discovery	✓	×	✓	×	×
Proposed Work	DAG Community Detection through Similarity Criteria	✓	✓	✓	✓	✓

[1] 1: Large Datasets; 2: Impact Analysis; 3: Multi Modal; 4: Filters Noise; 5: Outperforms Existing Models

The above works are compared in Table 1. As evident, the existing works employ several different strategies for influencer impact analysis and are also effective. However, they are computationally heavy due to the use of complex

models. Our proposed methodology is lightweight and is as effective if not more in a few of the cases.

3 Problem Statement and Methodology

The problem at hand is to determine the impact of influencers on social media users. To this end, we employ Directed Acyclic Graph (DAG) methodology for community detection and further perform impact analysis in subsequent subsections.

3.1 Mathematical Modeling

Let $D = \{ t_1, t_2, ..., t_n \}$, be the dataset, with t_i as individual data sample. Our objective is to identify similar communities within this dataset by constructing a network, leveraging attributes of the data (we consider three such attributes P, I,and E). These parameters are used to compute the similarity score (denoted by S), among the data points, based on which the communities are detected.

Initially, we construct a DAG, G, to represent the network structure of D. This DAG consists of nodes and edges between them. Edges represent the similarity between the nodes and thus are used to establish communities. To construct the edges, we iterate over each pair of data points in the D. If their similarity scores them exceeds a given threshold, we add a directed edge between them. Thus, DAG represents the communities with similar parameters and is further subjected to quality analysis to determine the impact of influencers on the communities identified.

3.2 Methodology

Our comprehensive workflow comprises three key phases as depicted in Figure 1. These phases are described below.

- *Data Generation* - In the first phase, we collect real-world data from X (Twitter) by harnessing web scraping techniques. This process involves retrieving tweets, user information, and its associated metadata. Upon obtaining the raw data, we perform data wrangling to extract specific features required for our analysis. This includes pre-processing steps like cleaning, feature extraction, and rearranging the data into a suitable structure for further processing.
- *Community detection* - The second phase involves community detection using our novel methodology based on the concept of DAGs. We use similarity scores to detect nodes with common features and identify them with the same communities. Also, we perform the same community detection using an ML model, to later compare the results.

– *Impact Analysis and Comparison* - In the final phase, we perform a comprehensive impact analysis and a thorough comparison between our work and the standard ML model, particularly the K-means clustering. We evaluate the models based on certain performance evaluation metrics. Through this, we demonstrate the superiority of our model in community detection and impact analysis of the influencers.

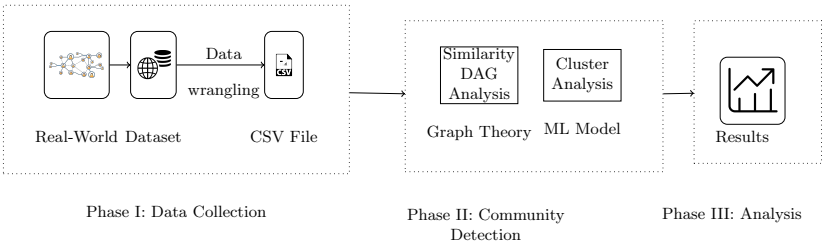


Fig. 1: Overall Workflow of the Proposed Work.

The computation of the similarity score involves assessing the resemblance between two nodes across various dimensions within the social network. Community detection in the context of SNA involves identifying groups of nodes that exhibit strong interactions within the network. Let $G = (V, e)$ represent the social network graph, where V denotes the set of vertices and e represents the set of edges (connections) between vertices. To formulate communities within this network, we define a similarity threshold T that determines the minimum similarity required for two vertices to be considered part of the same community.

We consider three types of vertices. Firstly, the product similarity P, determines if two users share common product preferences. Secondly, the influencer similarity I, to determine the impact of the influencer on the users, and lastly, engagement similarity E, that evaluates the similarity in interaction patterns between users and influencers. Using these three parameters we find the similarity score, S. To find the score, we first define the types of matches between the vertices u and v.

1. Ideal Match ($u \equiv v$): If the signature values of u and v are the same, then it is an ideal match. Here, nodes are of similar types.
2. Limited Match: In this type of match, the vertices are partially similar. We consider the following sub-types of the limited match:
   (a) Extension Match ($u \supset v$): If the signature value of u is greater than that of v, *i.e.*, u node is superior than v.
   (b) Encompass Match ($u \subset v$): When the signature value of v is greater than that of u, *i.e.*, u node is inferior to the v.
   (c) Approximate Match ($u \cap v$): When the signature value of v is a multiple of u's value, *i.e.*, u node supersedes the v node's features.

3. Unrelated Match ($u \neq v$): If the signature value of v is different from that of u, *i.e.*, nodes are dissimilar to each other.

We calculate the similarity score, S, using the below equation (p is the number of features matching between u and v).

$$S(u,v) = \sum_{i=1}^{p} \frac{F_{sim}(u,v)}{P} \tag{1}$$

Here, F_{sim} denotes the feature similarity between u and v. It is calculated as follows.

$$F_{sim} = \begin{cases} 1, & \text{if } \textit{ideal match} \\ 0, & \text{if } \textit{unrelated match} \\ \frac{\alpha}{\beta}, & \text{otherwise} \end{cases} \tag{2}$$

where, α is the balancing factor that controls the similarity score and thus the size and characteristics of the communities detected, while β holds the number of exact matching features between u and v. Algorithm 1 lists the steps required to calculate this score between two nodes.

Algorithm 1 Algorithm for Node Similarity Calculations.

Input: (1) Dataset, $D = \{t_1, t_2, ..., t_n\}$;
(2) Node Features, $F = \{P, I, E\}$;
Output: Node Similarities, S
$S \leftarrow \emptyset$;
for each $u, v \in D$ **do**
  for each $f \in F$ **do**
    $P \leftarrow$ number of matching features between u and v ;
    $F_{sim} \leftarrow$ calculate using Eq. 2 ;
    $s(u,v) \leftarrow$ calculate using Eq. 1 ;
    $S \leftarrow s(u,v) \cup S$;
  end for
end for
return S ;

In Algorithm 2, we present our methodology for community detection. Initially, an empty list of communities is created to store the identified communities. The algorithm iterates over each node u in the sorted dataset and selects it as the seed of a potential community. For each seed node, the algorithm traverses the remaining nodes to assess their similarity with the seed. If the similarity score exceeds the threshold T, the node is added to the community, and subsequently removed from the list of remaining nodes to avoid redundancy.

Computing the impact of influencer parameters involves analyzing various centrality measures, including degree centrality, closeness centrality, and betweenness centrality. We perform this analysis in the next section.

Algorithm 2 Community Detection Using DAGs

Input: (1) Dataset, $D = \{t_1, t_2, \ldots, t_n\}$;
    (2) Node Features, $F = \{P, I, E\}$;
    (3) Similarity Threshold, T;
Output: Communities, C
$S \leftarrow$ `calculate_similarity`(D, F) # using Algo. 1
$G \leftarrow$ `initialize_dag`(D, S, T)
$sorted\_nodes \leftarrow$ `topological_sort`(G) # linear ordering of nodes
$C \leftarrow \emptyset$
$visited \leftarrow \emptyset$
for each $u \in sorted\_nodes$ **do**
  if $u \notin visited$ **then**
    $community \leftarrow \{u\}$
    for each $v \in G$.`successors`(u) **do**
      if $S[u][v] \geq T$ **and** $v \notin visited$ **then**
        $community \leftarrow community \cup \{v\}$
        $visited \leftarrow visited \cup \{v\}$
      end if
    end for
    $C \leftarrow C \cup \{community\}$
  end if
end for
return C

4 Experiments and Results

In this Section, we present the implementation details, experiments performed, and results obtained. For our implementation, we used a machine with Intel Core i7 processors and up to 16 GB RAM. The proposed methodology was developed in Python, using packages NetworkX, scikit-learn, and pandas. Further, we constructed a Twitter dataset through the scrapping technique as there exist no standardized datasets for paid promotions analysis. Table 2 summarizes the details details of the collected data.

Table 2: Dataset Details

Attributes	**Value**
Number of Instances (Paid Promotion Tweets)	100,000
Number of Influencer	100
Number of Products	50
Number of Users	10,000

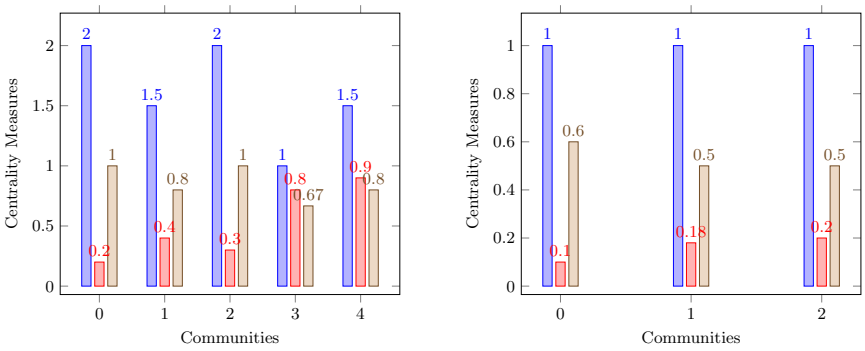
4.1 Experimental Setup

For the construction of DAG, we made assumptions. The similarity threshold's, T, value in the Algorithm 2 plays a crucial role in deciding the connections between nodes, and thus charting the formation of communities. We selected a threshold value of 0.5 through empirical experimentation, aiming to strike a balance between inclusivity and granularity, ensuring that communities are cohesive while avoiding excessive fragmentation.

In the experiments, we computed and analyzed the performance metrics. We scrutinized community sizes, centrality measures (*i.e.*, degree, betweenness, and closeness), and overall network coherence. Additionally, we undertook computation cost comparisons to evaluate the efficacy of the proposed method in community detection and impact analysis.

4.2 Results and Discussions

In this subsection, we present the results of two experiments. In the first experiment, we evaluate the centrality measure scores to determine the impact of the communities on paid promotions for the proposed methodology and compare it with the k-Means algorithm. Further, in the second experiment, we compare the computation time of both methods.



(a) Proposed Method (b) k-Means Clustering

Fig. 2: Centrality Measure Scores (– Degree Centrality, – Betweenness Centrality, – Closeness Centrality).

Comparison of Centrality Scores We evaluated the quality of communities identified by both methods using three metrics. We performed multiple experiments and averaged the results obtained, Fig. 2 depicts these results. Across multiple clusters, the DAG model consistently yields better centrality scores than k-Means. These elevated centrality scores signify that nodes within the

paid promotion tweets network are more cohesive and centrally positioned in the DAG model compared to the ML model. This phenomenon underscores the DAG model's ability to capture and leverage the inherent relationships and dependencies among users, tweets, and products within the network.

Comparison of time efficiency In this experiment, we evaluate the computational cost of both methods concerning the time taken for similarity score calculation in the proposed method *vs.* the distance calculation in the k-Means algorithm. As seen in Fig. 3, the DAG model consistently outperforms the ML model and demonstrates faster execution times. This is attributed to the fact that distance-based clustering algorithms make use of computationally intensive distance scores compared to our lightweight proposed similarity matches.

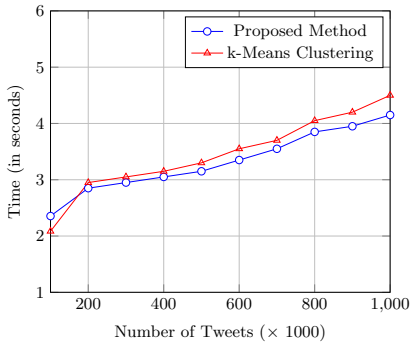


Fig. 3: Computation Time for Similarity *vs.* Distance Calculations.

Discussions Our proposed similarity score-based community detection approach outperforms the ML model in several key aspects, contributing to its superior performance. Firstly, the DAG model leverages the inherent structure of the data, capturing complex relationships and dependencies among nodes more effectively than ML algorithms. By representing the data as a graph, our approach encapsulates the intricate interconnections present in real-world networks, with greater fidelity. This allows for more nuanced analysis and interpretation of the data, leading to more accurate predictions and insights. As a result, our DAG-based approach not only outperforms ML models in terms of accuracy and speed but also offers enhanced interpretability and scalability, making it well-suited for paid promotion marketing analysis.

5 Conclusions

This work presented a promising algorithm for the detection of communities, showcasing the use the graph-theory-based approaches especially considering

DAGs. By harnessing the power of DAGs we developed a framework for similarity detection using our novel similarity scores which offers robust and efficient solutions for community detection tasks. The demonstrated performance gains showcase the applicability of our work in impact analysis of the influencer's paid promotion tweets. In the future, the integration of the proposed framework with additional features and refinement to diversify its domain and the applications in other fields where community detection plays a crucial role can be explored.

References

1. Arora, A., Bansal, S., Kandpal, C., Aswani, R., Dwivedi, Y.: Measuring Social Media Influencer Index- insights from Facebook, Twitter and Instagram. Journal of Retailing and Consumer Services **49**, 86–101 (2019). https://doi.org/10.1016/j.jretconser.2019.03.012
2. Chikhaoui, B., Chiazzaro, M., Wang, S.: Discovering and Tracking Influencer-influence Relationships between Online Communities. pp. 1–9 (10 2015). https://doi.org/10.1109/DSAA.2015.7344846
3. Dhun, Dangi, H.K.: Influencer Marketing: Role of Influencer Credibility and Congruence on Brand Attitude and eWOM. Journal of Internet Commerce **22**(sup1), S28–S72 (2023). https://doi.org/10.1080/15332861.2022.2125220
4. Khanam, K.Z., Srivastava, G., Mago, V.: The Homophily Principle in Social Network Analysis: A survey. Multimedia Tools and Applications **82**(6), 8811–8854 (Mar 2023). https://doi.org/10.1007/s11042-021-11857-1
5. Kilipiri, E., Papaioannou, E., Kotzaivazoglou, I.: Social Media and Influencer Marketing for Promoting Sustainable Tourism Destinations: The Instagram Case. Sustainability **15**(8) (2023). https://doi.org/10.3390/su15086374
6. Li, M., Lu, S., Zhang, L., Zhang, Y., Zhang, B.: A Community Detection Method for Social Network Based on Community Embedding. IEEE Transactions on Computational Social Systems **8**(2), 308–318 (2021). https://doi.org/10.1109/TCSS.2021.3050397
7. Logan, Austin P. and LaCasse, Phillip M. and Lunday, Brian J.: Social Network Analysis of Twitter Interactions: A Directed Multilayer Network Approach. Social Network Analysis and Mining **13**(1) (2023). https://doi.org/10.1007/s13278-023-01063-2
8. Rashid, Y., Bhat, J.I.: Topological to Deep Learning Era for Identifying Influencers in Online Social Networks :a Systematic Review. Multimedia Tools and Applications **83**(5), 14671–14714 (Feb 2024). https://doi.org/10.1007/s11042-023-16002-8
9. Ríos, S., Aguilera, F., Núñez-González, J., Graña, M.: Semantically Enhanced Network Analysis for Influencer Identification in Online Social Networks. Neurocomputing **326-327** (2017). https://doi.org/10.1016/j.neucom.2017.01.123
10. Zengin Alp, Z., Gunduz Oguducu, S.: Identifying Topical Influencers On Twitter Based on User Behavior and Network Topology. Knowledge-Based Systems **141** (2017). https://doi.org/10.1016/j.knosys.2017.11.021
11. Zheng, C., Zhang, Q., Young, S., Wang, W.: On-demand Influencer Discovery on Social Media. Proceedings of the 29th ACM International Conference on Information & Knowledge Management p. 2337–2340 (2020). https://doi.org/10.1145/3340531.3412134