

A Framework for Identifying Influential Spreaders of Drug Use in Online Social Networks

Reem Al-Ghannam

Information Systems Department
College of Computer and Information
Sciences
King Saud University
Riyadh, Saudi Arabia
r.g.alghannam@gmail.com

Mourad Ykhlef

Information Systems Department
College of Computer and Information
Sciences
King Saud University
Riyadh, Saudi Arabia
ykhlef@ksu.edu.sa

Hmood Al-Dossari

Information Systems Department
College of Computer and Information
Sciences
King Saud University
Riyadh, Saudi Arabia
hzaldossari@ksu.edu.sa

Abstract— Online social networks have become key platforms for spreading drug-related content, which presents a risk to the public, especially for people who are more susceptible to drug use and addiction. Generating vast amounts of data, such networks offer valuable insights for computational analysis. In this paper, we introduce a new framework to identify influential spreaders who use social networks to promote drug-related content. It does so by means of influence maximization (IM) algorithms; in particular, it uses an approach called ‘label neighbor-based selection of seed nodes (LNSSN)’. It also proposes new metrics for the percentages of types of users affected, comprehensively assessing each algorithm’s effectiveness across different network types and combinations of scenarios. This framework is designed to assist governments, technology companies, and law enforcement agencies in prioritizing and addressing sources of drug-related content on social networks.

Keywords—drug use, influential spreaders, influence maximization, information diffusion, social network analysis

I. INTRODUCTION

Online social networks (OSNs) have transformed global communication by making remote, instant interactions possible. However, drawbacks have followed from this [1]. In early 2024, the number of social media users was reported to be approximately 5.07 billion people, and 259 MILLIONS of these had joined within the previous year. About 62.6% of the world’s population access major platforms like X (formerly Twitter), Facebook, and YouTube, using them, on average, for 2 hours and 20 minutes per day [2]. These platforms have become essential in daily life, offering numerous benefits; however, this vast number of users makes effective monitoring of user-generated content increasingly difficult. At the same time, competition among these networks creates pressure for innovation so as to attract and retain user engagement.

There is general acceptance that social media networks can enable activities that are undesirable or illegal. For instance, users may exploit features such as direct messaging to coordinate criminal activities, including drug distribution. Furthermore, many studies have identified a strong correlation between excessive social media usage and youth drug addiction [3]. More generally, research shows that when people, especially young people, are exposed to alcohol-related content on social media, they are more likely to consume alcohol, use drugs and eventually become addicted.

The World Health Organization (WHO) and the United Nations Office on Drugs and Crime (UNODC) both operate

constitutionally to tackle drug use and addiction. The WHO works to reduce drug demand by preventing drug use and managing drug use disorders. To do so, it collects, analyses and disseminates evidence-based policies aimed at blocking access to drugs, early intervention when a risk of addiction is suspected, and both treating and rehabilitating addicts. The WHO also supports monitoring efforts and the mitigation of drug use effects through relevant information and technical assistance at all levels [4]. Alongside these efforts, Saudi Arabia’s General Directorate of Narcotics Control (GDNC) tackles drug-related issues offline, such as the real-world trade in illicit drugs, and online, combating drug-related activities on the internet. With social media platforms evolving rapidly, those engaged in illegal drug dealing have quickly adapted ways of reaching and engaging with young people [5].

Criminal groups often exploit online social networks (OSNs) to remain anonymous while promoting illicit activities, and extending their reach to drug users globally. By its sheer volume, social media traffic increasingly complicates the detection and elimination of criminal content, including that shared by drug dealers and other illegal operators. Most online platforms employ traditional manual methods of moderation, which are often overwhelmed. Consequently, governments, anti-drug agencies, cybersecurity firms, and other stakeholders are concerned to develop efficient and fully automated systems to identify drug-related content online and combat its spread.

This background explains the importance of the proposed study, which uses advanced computational approaches to target the identification of key influencers on social media platforms. These methods include a highly effective, state-of-the-art subset of techniques for information diffusion [6], and influence maximization (IM) techniques [7]. Despite the development of a promising social network analysis (SNA) application utilizing IM techniques to tackle this task, the inherent complexity of the problem remains widely recognized and worthy of continued research in this area. It is particularly difficult to detect drug-related references on platforms like X because the messages are extremely short, which often limits the ability to clearly establish the meaning in context.

Researchers from various fields, including computer science, social sciences, and psychology, have explored a range of approaches to detecting media users spreading drug-related material online. In most studies, interdisciplinary teams focus either on analyzing online user behavior or developing practical methods for quickly identifying individuals using the internet to engage in drug-related activities [8]. The identification of influential users—those

with a significant impact on others—is a key area of social network analysis. This paper suggests a framework for identifying influential users who spread drug-related content on X, helping to tackle this urgent issue more effectively.

This paper is structured as follows: Section 2 provides details of the proposed framework, divided into four key parts. Network dataset representation, including dataset construction and visualization preparation, is covered in Section 2.1. Information diffusion models are explored in Section 2.2. Section 2.3 addresses influence maximization (IM), while a strategy for selecting seed nodes for IM algorithms is addressed in Section 2.4. Section 3 provides a step-by-step overview of the complete framework, and Section 4 concludes the paper.

II. FRAMEWORK

A. Representation of the Network Dataset

A social network can be depicted as a weighted-directed graph, often termed an influence graph, denoted as $G = (V, E, W)$, comprising N nodes (users) and M edges. In this representation, V represents the set of users within the network, while E signifies a set of relationships, such as friendships, followers, re-posts, mentions, replies, co-hashtags, co-authorship, etc. It is important to note that the nature of these relationships depends on the specific characteristics of the network. The edge-weight, denoted as W , reflects the strength of the relationship between individuals [9].

Based on the definition of the social network, nodes within the social network are used to construct the network. In this study, X networks are utilized to verify the effectiveness of the proposed method for identifying influential spreaders of information compared to existing approaches. X is a widely used OSN through which millions of users can broadcast short messages to social connections, providing a natural environment for studying diffusion processes. Unlike other OSNs, X is specifically designed for information dissemination, with users following the broadcasts of others, enabling information diffusion networks to be reconstructed by analyzing the follower network.

The metrics extracted from OSNs can be crucial to the identification of influential users, but different metrics used in analysis can lead to variation in rankings, so it is essential to understand how these metrics correlate with users identified as influential. There are two categories of metrics used in such analyses: network-based and post-based [10].

Network-based metrics focus on the connections and interactions among users within the network. These metrics are tied to the structure of the network. For instance, the in-degree metric reflects the size of a user's audience, while the propagation metric shows how information spreads through the network, indicating a user's ability to generate and distribute content. In contrast, engagement metrics measure a user's tendency to involve others in conversations. On platforms like Facebook and X, in-degree metrics are represented by the number of 'friends' and 'followers', respectively. Propagation is measured by 'sharing' on Facebook and 'reposting' on X. Engagement is captured by 'tagging' on Facebook and 'mentioning' on X.

Post-based metrics, on the other hand, assess the quality of content posted by a user. These metrics focus on the features that make a post go viral; an influential user's ranking may be enhanced when other users post similar content.

X has a data policy that permits anonymized network datasets to be shared, including user relationships: this means that publicly available networks can be used for evaluation. This study uses three real-world X networks to assess the improved method's performance including both directed and non-directed networks. As shown in figure 1, the 'reply', 'mention' and 'hashtag' networks were used to extract the number of mentions, replies, and co-hashtags associated with each user.

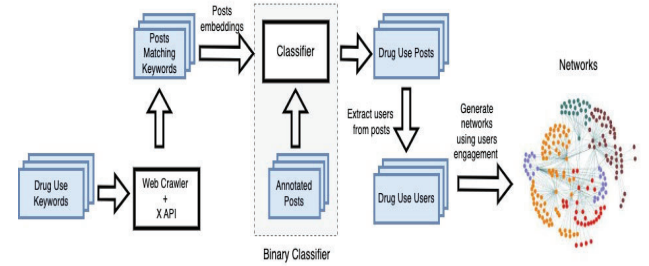


Fig. 1. Steps of networks formation [11].

The three networks – mentions, replies and co-hashtags – were merged in order to analyze the combined effect of different types of interactions. Various combinations of the merged networks were tried to explore the contribution of different interaction types to information diffusion. Table 1 below summarizes the key structural properties of these networks.

TABLE 1. DETAILS OF EACH NETWORK

Networks	Nodes	Edges	Average Clustering
Hashtags Network (HT)	3728	323676	0.8085
Replies Network (RN)	39093	22482	0.0010
Mentions Network (MN)	6125	4596	0.0109
Network RN, MN	41519	24745	0.0048
Network HT, MN	9646	328244	0.3180
Network RN, HT	42261	345531	0.0715
Merged HT, RN, MN	44583	348381	0.0706

The analysis shows significant variation in the structural properties of the networks (as shown in figures 2, 3, and 4) depending on the type of interaction they represent. Communities around hashtags tend to be tightly-knit, while replies show greater dispersion. With moderate clustering, mention networks show a structure between the other two. An understanding of these structural differences is crucial when selecting appropriate seed nodes for models of the spread of infection. In particular, the clustering coefficient

can help to identify the expected spread dynamics, as greater clustering suggests more localized spread.

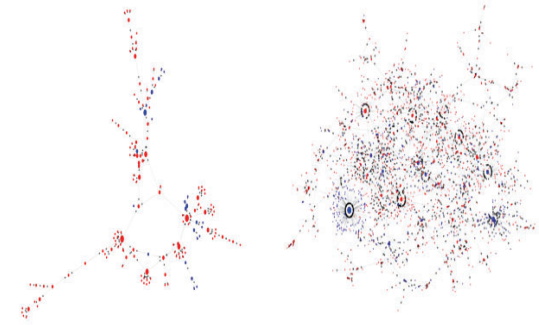


Fig. 2. Mentions network graph. Fig. 3. Replies network graph.

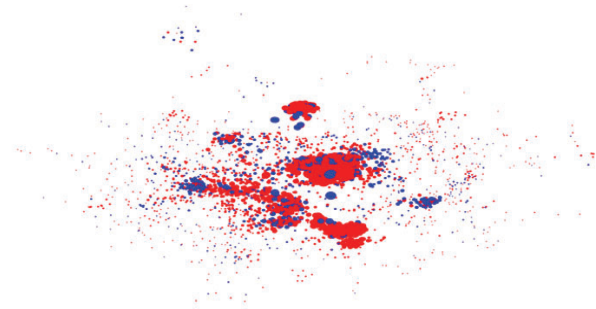


Fig. 4. Co-Hashtags network graph.

In all of the above networks from the dataset, 60% of posts have 60% T values from the target feature (drug-related posts) and 40% of F values (non-drug-related posts). On this basis, the first part of our framework will focus on the network dataset representation of the online social network.

B. Information Diffusion Models

Information diffusion occurs when information is passed between individuals or communities in a network. This is also known as information propagation or information spread. Pieces of information can be exchanged or diffused between individuals using social networks, and the diffusion process enables a substantial amount of information to be mined [12]. Over the past decade, the diffusion process has been extensively analyzed for rumor control, market predictions, and the detection of cybercrime and/or misbehavior. Social networks have developed the diffusion models that view users as subject to influence from the actions of their neighbors; thus, in social networks, the diffusion process occurs through cascades of information. These cascades can be conceptualized as directed trees, in which the first node of the activation sequence is the tree's root node [13]. Examining the tree enables us to determine the pattern of interactions and influences that characterize the diffusion of information.

Applications of diffusion models range from detecting breaking news to strategic marketing, thanks to their versatility in accounting for and simulating the dissemination of information over social networks. Based on the analysis of data, the authors in [14] divided current research on modeling information diffusion into two categories: theory-centered and data-centered. Theory centered models are principally used in economics, epidemiology, and sociology; the independent cascade model, linear threshold model, and epidemic model, with their extensions, are well known and

have been extensively investigated. Although theory centered models provide reasonable, albeit partial, solutions to the problem of information diffusion, their limitations must be recognized. One such limitation of these models is that they typically use parameters that are randomly distributed rather than learned from real-world diffusion data. Furthermore, theory-centered models cannot operate in real-time. Because of these limitations, theory-centered models cannot generate insights into the procedural aspects of information diffusion.

Unlike theory-centered models, data-centered models are normally trained on real-world data concerning information diffusion. There are two broad sub-categories: macro-models and micro-models [15]. The former, are also known as cascade generation models, can generate cascades of information diffusion, whose macro-level properties are similar to those typical of real-world diffusion cascades. For use in making predictions, however, this kind of model suffers from the same limitations noted in relation to theory-centered models. Micro-models are important precisely because they address this deficiency by the ability to predict whether certain information will activate a given social network user [16]. In [17], the researchers demonstrated a real-time information diffusion model for X, on the basis of network parameters that included centrality, density, degree, clustering coefficient, average path length, and diameter. Further, Kumar et al. [18] developed a diffusion model based on models (Bass and SIS) which are principally used to examine epidemiological trends in large networks by simulation. For rumor detection, Hu et al. [19] designed the Susceptible-Hesitating-Affected-Resistant (SHAR) model, with the aim of considering different individuals' attitudes to the dissemination of rumors.

The literature includes various diffusion models through which researchers have attempted to understand the processes involved in information diffusion. The two most prevalent diffusion models are the independent cascade model (IC) and the linear threshold model (LT). The authors of [20] proposed two models for using historical cascades to interpret person-to-person influence in information diffusion. Their first model employs the linear threshold propagation approach, while the second uses random walk ideas to examine multi-step diffusion of influence between user pairs. These models are trained by means of mixed integer programs and approximate methods in artificial neural networks. Jayalath et al. [21] later used complex networks to explore information diffusion, analyzing several agent-based models on directed scale-free networks. Their models included linear threshold (LT) models, independent cascade (IC) models, and stochastic linear absolute threshold (SLATM) models. With these, they simulated the adoption and spread of information over time, so as to understand whether conceptually different models can represent the same underlying generative structures. The paper identified distinctive properties in the capabilities of each model.

As exemplified in the efforts discussed above, information propagation has attracted significant attention. However, because it is a complex process involving many factors, it is important for models to consider a wide range of influencing factors to approximate to the dynamics of real-world communication.

C. Influence Maximization

Influence maximization (IM), originating in viral marketing, is seen as an optimization problem. Its goal is the identification of an initial group of seed users to maximize the spread of influence throughout the network [22], computing the influence propagation of each active node based on the diffusion model.

The process of influence maximization begins with identifying a set of influential users, termed a seed set and represented by $S \subseteq V$, where $k < |V|$ nodes, within a social network $G = (V, E)$ [23]. To do so, the number of influenced users in the network G is maximized by propagating a diffusion model, where V is the set of nodes (users) in G , and E is the set of edges (both directed and undirected) which represents social connections between users. The problem can be expressed mathematically as:

$$IM_M(G, k) = \underset{e \subseteq V, |e|=k}{\operatorname{argmax}} \sigma_M(e, G) \quad (1)$$

Where the function σ calculates the spread of influence achieved by a given set of nodes when activated in e . The objective is to identify the subset e of size k within V that leads to the maximum spread of influence in the social network [22]. The problem of influence maximization (IM) is known to be NP-hard, meaning that it is not practical to obtain the optimal solution. Thus, it is possible to achieve an approximate solution if and only if the expected function of influence spread, represented as σ_M , submodular; this is the case if it is both monotone increasing and demonstrates diminishing returns [23]. These properties are explained as follows:

Property 1: Monotone Increasing

An objective function $\sigma(S)$ is termed ‘monotonically increasing’ if $\sigma(S) \leq \sigma(T)$, where $S \subset T$.

Property 2: Diminishing Return

An objective function $\sigma(S)$ exhibits diminishing returns if $\sigma(S \cup u) - \sigma(S) \geq \sigma(T \cup u) - \sigma(T)$, for all $u \in T$ and $S \subset T$.

To address the IM problem, an initial assessment is required of the propagation influence, represented by $\sigma(S)$, for a specified seed set S , after which the goal is to maximize this influence. However, there are theoretical challenges to both of these stages. Estimating the propagation influence $\sigma(S)$ of a given seed set S faces theoretical complexity which is referred as the propagation influence problem. Similarly, maximizing the propagation influence faces the theoretical difficulty referred to as the IM problem [23].

The authors of [24] demonstrated that the theoretical difficulty of gauging the spread of influence $\sigma(S)$ within a specified seed set S is #p-hard. In classical diffusion models, computing the propagation influence $\sigma(S)$ for a given set S of activated users is similarly shown in [25] to be a problem of #p-hard complexity. In accordance with the definitions above, it is implausible to generate an optimal solution within polynomial time unless $P = NP$.

The IM problem is NP-hard because of the theoretical difficulties inherent in maximizing the propagation influence $\sigma(S)$. This has been shown for classical diffusion models in [22]. In general, the influence maximization problem is challenging because it is computationally complex,

compounded by the size of networks, the difficulty of finding optimal solutions, uncertainties around influence propagation, and limited knowledge of network structure. These challenges continue to be explored through various approximation algorithms and heuristics with the aim of deriving effective solutions for real-world instances.

Influence maximization models aim to identify a small set of influential spreaders with maximal ability to spread information across a network, for which several techniques have been proposed. The IM problem has been widely researched across multiple disciplines because it has a wide range of applications, including maximizing information spread, identifying the most influential nodes [26], preventing the spread of misinformation [27], and blocking the propagation of negative influences, such as rumors and misinformation in social networks [28], as well as computer viruses [29]. In practice, IM is defined as the challenge of identifying a small set comprising the most influential nodes which are active in a social network. As well as the regulatory functions of limiting misinformation and malware, this task is crucial for effective advertising and promotion on social media platforms [30]. Once the information diffusion model is defined, selecting the optimal set of seed nodes allows information to spread to the maximum possible extent. Two types of algorithms are commonly employed to address this challenge: greedy algorithms and heuristic algorithms [31].

Kempe et al [7] introduced the greedy algorithm, which essentially works by iteratively selecting the most influential node as a seed node. In each iteration, multiple Monte Carlo simulations serve to estimate the marginal influence spread of each node. This algorithm has the advantage of efficiently approximating the solution for optimal influence maximization, ensuring that the propagation strategies are effective [32]. However, this is an NP-hard problem, and the greedy algorithm is highly computationally complex. This has led to the development of several alternative algorithms for the influence maximization problem [33]. Leskovec et al.’s cost-effective lazy forward (CELf) algorithm [34] is an optimized form of the greedy algorithm that utilizes lazy forward evaluation to minimize the number of simulations needed to calculate information spread: this improves the speed of influence computation by 700 times. The CELf++ algorithm further enhances the efficiency of CELf by implementing a backward pruning technique. These algorithms calculate the influence spread only once for each node, and then identify which node is the most influential, after which the lazy evaluation process is applied [35].

In turn, Aghaee et al. [36] addressed the IM problem through the ‘group of influential nodes’ (GIN) algorithm, based on the greedy method. In comparison with other algorithms across seven real-world datasets, this cut down the number of simulations, made the proposed approach faster, simplified the search to identify influential users, and achieved a high level of influence spread with the lowest runtime.

Zhu et al. [37] designed the ‘dominating influence maximization’ (DomIM) algorithm to find a local optimum when applied to the IM problem. DomIM proved to be the most efficient and reliable of the approaches tested, generating optimal results on three network topologies using different diffusion models. For an enhanced solution during the local search phase, the influence of each newly

constructed seed set was calculated in every iteration. However, this process may slightly affect the overall efficiency of the proposed method.

Another drawback is the scalability challenge of the greedy algorithm, which significantly limits its application to large-scale online social networks and makes it less suitable for widespread propagation efforts. Heuristics offer an alternative approach, various heuristics having been introduced by some researchers to simplify the diffusion process [38]. For example, Ko et al. [39] proposed combining path-based and community-based strategies to create a hybrid model, while Kim et al. [40] used a random walk and rank merge in an effective pruning technique. Additionally, with high-centrality nodes often being the most "important" in a network, a simple heuristic alternative to the greedy algorithm can be developed using centrality measures to estimate node influence, selecting the seed set according to centrality rankings. Common centrality measures include degree centrality [41], betweenness [42], closeness [32], and k-shell (or k-core) [33]. Nevertheless, despite their computational efficiency, the performance of heuristics can vary significantly across different networks. Hence, these algorithms may perform well in some networks, but yield poor results in others. Consequently, a need remains for a solution that is both effective and efficient across a variety of network types.

It is important to note that most prior studies have sought to identify influential spreaders by means of either greedy or heuristic approaches, but measurement algorithms should be chosen according to the application that is needed. For instance, if time efficiency is mission-critical but quality is less essential (e.g., for disabling X accounts that promote drug use), a greedy algorithm, being more computationally intensive, will be less effective than a heuristic algorithm.

D. Identifying Seed Nodes

In any network diffusion model to check the spread of infection or information, the first stage is the selection of the optimal set of seed nodes. The diffusion process starts from these seed nodes, so they must be selected strategically to maximize the spread of influence. The main approach to seed selection in this method is based on degree centrality, a common measure of the importance of nodes in a network. The method will be evaluated based on extensive experiments on real-world datasets to compare the proposed methods, using the heuristic, greedy, and CELF++ algorithms as baselines underlying independent cascade (IC) and linear threshold (LT) models [34], [35].

The selection of seed nodes that are most likely to maximize influence spread will leverage centrality-based seed selection based on their degree centrality and a specific criterion ("label" = "T"), without considering their neighbors' characteristics. To identify the most influential seed nodes, the degree centrality will first be calculated for all nodes in the graph and then they will be sorted accordingly. Next, selection criteria will be applied to identify nodes that meet the "label" = "T" condition. The top nodes will finally be selected based on the chosen strategy. The selected nodes are expected to be those most likely to spread influence widely throughout the network because of their direct connections, i.e., their degree centrality. Figure 5 presents an overview of the network seed node identification process. The candidates will be generated

in the input network and then the seed selection procedure will be applied to get the seed nodes.

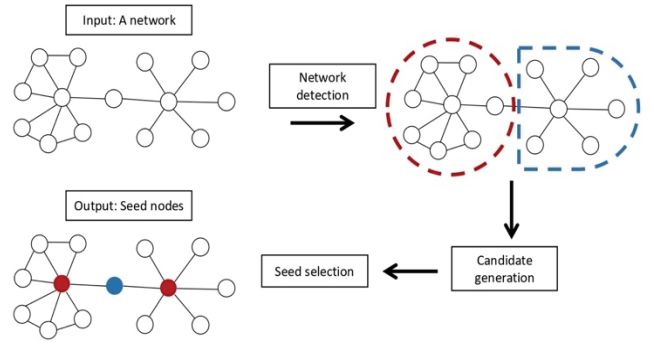


Fig. 5. Identifying seed nodes in the network.

III. DISCUSSION

We will use our framework to evaluate network infection dynamics using the independent cascade (IC) and linear threshold (LT) models, focusing seed node selection on three strategies: the heuristic approach, selecting nodes based on degree centrality; top degree centrality without neighbor consideration, which prioritizes nodes with high centrality; and the greedy and CELF++ approaches, which identify seed nodes that maximize infection spread. We will also enhance our approach to selecting seed nodes by considering a significant number of "T-labeled" neighbors connected to them—where a node labeled "T" means user-posted drug-use content—aiming to improve the efficiency of infection spreading through label awareness. In addition, we will introduce new metrics to assess the percentage of "T" and "F" nodes infected, so as to give a detailed view of each algorithm's performance across various network types and combined scenarios. Figure 6 illustrates the experimental process for developing and evaluating an effective method of identifying influential spreaders.

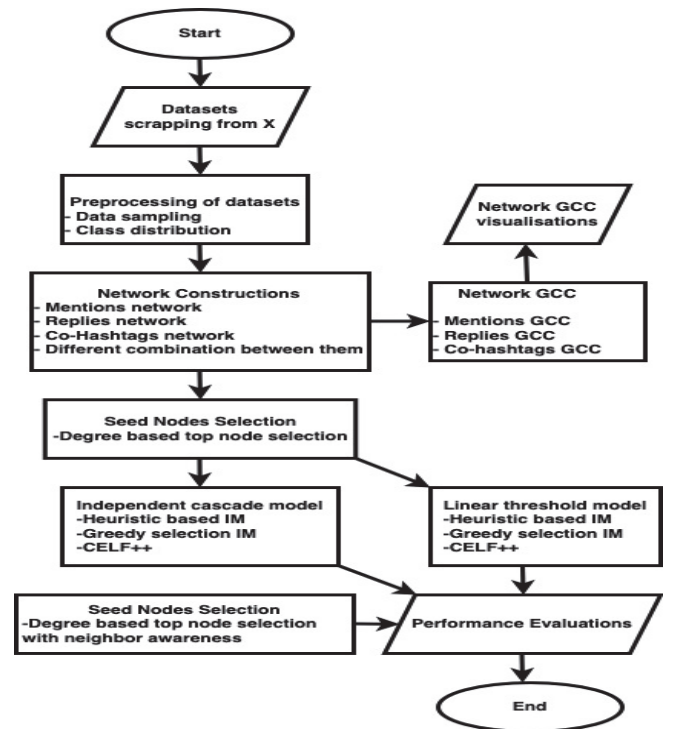


Fig. 6. Flowchart of the methodology for identifying influential spreaders.

Once the seed nodes are selected, the next step is to simulate how information spreads through the network. This process helps in understanding the effectiveness of the seed nodes and in measuring the potential reach of the influence. To simulate the spread of influence through a network based on a probabilistic model, where nodes have a certain chance of influencing their neighbors. The independent cascades (IC) model is a widely used model in network science for simulating information diffusion. In this model, an initially activated node (seed node) has a probability of activating its neighbors. This process will continue iteratively until no further activations occur. Also, we will implement the LT model for information diffusion in a network. It will simulate how influence spreads from a set of initial seed nodes across the network. The simulations will run on different network graphs, using both heuristic-based and optimized seed selection methods.

The procedure will begin with initializing the model by activating the selected seed nodes. The IC model will then run on the graph for a predefined number of iterations (nitr). During each iteration, the activated nodes will attempt to activate their neighbors, with the success of activation determined by a probabilistic rule. This process will continue until no new nodes are activated. After the simulation, the results will be processed to count the number of nodes in different states—infected, susceptible, and removed—and the extent of influence will be measured by the number of infected nodes across iterations. The expected outcome of this simulation will be to gain insights into the potential spread of influence initiated by the selected seed nodes, allowing for the evaluation of different seed node selection strategies.

Influence maximization models are superior to other influential spreader identification methods because they focus on selecting the optimal set of initial nodes (seed nodes) that can maximize the spread of influence across a network. These models account for the diffusion process and interactions within the network, making them more effective in identifying key influencers compared to methods that rely solely on static metrics (like degree centrality). Additionally, influence maximization approaches are specifically designed to target spread, which is the core objective in many real-world applications, such as viral marketing or information dissemination.

While the IC model provides a probabilistic framework for influence spread, the next step involves optimizing the selection of seed nodes using more sophisticated algorithms to further maximize influence. To iteratively select nodes that will maximize the spread of influence when added to the current set of seed nodes. The greedy algorithm is a heuristic method that selects nodes based on their marginal contribution to the overall influence spread. The procedure will begin with an empty set of selected nodes, and at each iteration, the node that provides the greatest increase in influence spread will be added. This process will continue until k nodes are selected. The expected outcome of the greedy algorithm will provide a near-optimal set of seed nodes that maximizes the spread of influence, effectively balancing computational efficiency with effectiveness.

Moreover, we will implement the CELF++ algorithm to enhance the efficiency of the greedy algorithm through optimization techniques, specifically aiming to reduce

redundant calculations. The CELF++ algorithm, an optimized version of the greedy algorithm, improved computational efficiency by minimizing the number of influence spread evaluations required. The procedure will involve calculating the initial marginal gain for each node, selecting nodes with the highest marginal gain, and updating and pruning candidate nodes as necessary. This process will continue until k nodes are selected. The expected outcome will be that the CELF++ algorithm would yield a set of seed nodes similar to those selected by the greedy algorithm but with significantly reduced computational overhead, making it more suitable for large-scale networks.

Top degree centrality with neighbor consideration selects nodes with the highest degree centrality that also have a large number of neighbors sharing a specific characteristic (in this case, "label" == "T"). This selection prioritizes nodes whose influence can be amplified by their immediate neighbors in the network. In the updated approach called label-neighbor-based selection of seed nodes (LNSSN), the method for selecting seed nodes will be refined to enhance the optimization of infection spread. Thus, nodes are selected based on their degree centrality with a focus on the number of "T" neighbors they have.

After running the simulation using different seed selection strategies, the performance will be evaluated by analyzing the number of nodes infected (influenced) during the simulation. The `get_iteration_results` function will extract the number of infected, susceptible, and removed nodes at each iteration. The maximum number of nodes infected will be used as the primary metric for evaluating the effectiveness of the seed selection strategy.

Such approaches to challenges should be carefully considered when testing the framework in real-world scenarios involving online social networks and selecting the appropriate techniques.

IV. CONCLUSION AND FUTURE WORK

This paper introduces a novel framework to identify seed nodes of drug use content using influence maximization algorithms with diffusion models. This unique approach has not previously been applied for identifying influential spreaders of drug use discussions in OSNs by strategically incorporating contextual features such as labels. To evaluate the effectiveness of the enhanced approach in assessing the spreading capability of nodes, new metrics will be proposed to provide a more comprehensive evaluation of algorithm performance.

In future work, the framework will be implemented using high-performance programming languages, network analysis techniques, information diffusion models, and influence maximization algorithms. Additionally, it will be tested on an online social network.

REFERENCES

- [1] T. Alsulimani, "Social Media and Drug Smuggling in Saudi Arabia," *Journal of Civil & Legal Sciences*, vol. 07, no. 02, 2018, doi: 10.4172/2169-0170.1000249.
- [2] D. Chaffey, "Global social media statistics research summary," Smart Insights.
- [3] R. Prieto Curiel, S. Cresci, C. I. Muntean, and S. R. Bishop, "Crime and its fear in social media," *Palgrave Commun*, vol. 6, no. 1, Dec. 2020, doi: 10.1057/s41599-020-0430-7.

- [4] "UNODC/WHO Program on Drug Dependence Treatment and Care," World Health Organization. Accessed: Jul. 07, 2023. [Online]. Available: <https://www.who.int/initiatives/joint-unodc-who-programme-on-drug-dependence-treatment-and-care>
- [5] A. AlSayyari and F. AlBuhairan, "Relationship of media exposure to substance use among adolescents in Saudi Arabia: Results from a national study," *Drug Alcohol Depend.*, vol. 191, pp. 174–180, Oct. 2018, doi: 10.1016/j.drugalcdep.2018.01.025.
- [6] F. Firdaniza, B. N. Ruchjana, D. Chaerani, and J. Radianti, "Information Diffusion Model in Twitter: A Systematic Literature Review," *Information (Switzerland)*, vol. 13, no. 1, Jan. 2022, doi: 10.3390/info13010013.
- [7] H. Li, S. Yang, and M. Xu, "Influence Maximization in Social Networks: A Survey."
- [8] Y. Fan, Y. Zhang, Y. Ye, X. Li, and W. Zheng, "Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from Twitter and case studies," in *International Conference on Information and Knowledge Management, Proceedings, Association for Computing Machinery*, Nov. 2017, pp. 1259–1267. doi: 10.1145/3132847.3132857.
- [9] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "Influence maximization diffusion models based on engagement and activeness on instagram," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2831–2839, Jun. 2022, doi: 10.1016/j.jksuci.2020.09.012.
- [10] M. Ali and D. Al-Garadi, "EFFECTIVE METHODS TO DETECT CYBERBULLYING AND INFLUENTIAL SPREADERS IN AN ONLINE SOCIAL NETWORK."
- [11] R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, "A Large-Scale Study of the Twitter Follower Network to Characterize the Spread of Prescription Drug Abuse Tweets," *IEEE Trans Comput Soc Syst*, vol. 6, no. 6, pp. 1232–1244, Dec. 2019, doi: 10.1109/TCSS.2019.2943238.
- [12] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," Sep. 29, 2017, MDPI AG. doi: 10.3390/info8040118.
- [13] R. M. Safari, A. M. Rahmani, and S. H. Alizadeh, "User behavior mining on social media: a systematic literature review," *Multimed Tools Appl*, vol. 78, no. 23, pp. 33747–33804, Dec. 2019, doi: 10.1007/s11042-019-08046-6.
- [14] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, "Modeling Information Diffusion over Social Networks for Temporal Dynamic Prediction," *IEEE Trans Knowl Data Eng*, vol. 29, no. 9, pp. 1985–1997, Sep. 2017, doi: 10.1109/TKDE.2017.2702162.
- [15] A. Iamnitich, L. O. Hall, S. Horawalavithana, F. Mubang, K. W. Ng, and J. Skvoretz, "Modeling information diffusion in social media: data-driven observations," *Front Big Data*, vol. 6, 2023, doi: 10.3389/fdata.2023.1135191.
- [16] J. Qi, X. Liang, Y. Wang, and H. Cheng, "Discrete time information diffusion in online social networks: micro and macro perspectives," *Sci Rep*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-018-29733-8.
- [17] Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies (NGCT): October 14th-16th, 2016, Center for Information Technology, University of Petroleum and Energy Studies, Dehradun. IEEE, 2016.
- [18] P. Kumar and A. Sinha, "Information diffusion modeling and analysis for socially interacting networks," *Soc Netw Anal Min*, vol. 11, no. 1, Dec. 2021, doi: 10.1007/s13278-020-00719-7.
- [19] Y. Hu, Q. Pan, W. Hou, and M. He, "Rumor spreading model with the different attitudes towards rumors," *Physica A: Statistical Mechanics and its Applications*, vol. 502, pp. 331–344, Jul. 2018, doi: 10.1016/j.physa.2018.02.096.
- [20] Z. Qiang, E. L. Pasiliao, and Q. P. Zheng, "Model-based learning of information diffusion in social media networks," *Appl Netw Sci*, vol. 4, no. 1, Dec. 2019, doi: 10.1007/s41109-019-0215-3.
- [21] C. Jayalath, C. Gunaratne, W. Rand, C. Seneviratne, and I. Garibay, "A GENERALIZATION OF THRESHOLD-BASED AND PROBABILITY-BASED MODELS OF INFORMATION DIFFUSION," *Adv Complex Syst*, vol. 26, no. 2, Mar. 2023, doi: 10.1142/S0219525923500054.
- [22] S. S. Singh, D. Srivastva, M. Verma, and J. Singh, "Influence maximization frameworks, performance, challenges and directions on social network: A theoretical study," 2021, King Saud bin Abdulaziz University. doi: 10.1016/j.jksuci.2021.08.009.
- [23] Y. Ye, Y. Chen, and W. Han, "Influence maximization in social networks: Theories, methods and challenges," Dec. 01, 2022, Elsevier B.V. doi: 10.1016/j.array.2022.100264.
- [24] S. Pei, J. Wang, F. Morone, and H. A. Makse, "Influencer identification in dynamical complex systems," 2020, Oxford University Press. doi: 10.1093/commnet/cnz029.
- [25] G. Tong, R. Wang, and Z. Dong, "On Multi-Cascade Influence Maximization: Model, Hardness and Algorithmic Framework," *IEEE Trans Netw Sci Eng*, vol. 8, no. 2, pp. 1600–1613, Apr. 2021, doi: 10.1109/TNSE.2021.3065272.
- [26] H. Li, R. Zhang, Z. Zhao, X. Liu, and Y. Yuan, "Identification of top-k influential nodes based on discrete crow search algorithm optimization for influence maximization," *Applied Intelligence*, vol. 51, no. 11, pp. 7749–7765, Nov. 2021, doi: 10.1007/s10489-021-02283-9.
- [27] P. Wu and L. Pan, "Scalable influence blocking maximization in social networks under competitive independent cascade models," *Computer Networks*, vol. 123, pp. 38–50, Aug. 2017, doi: 10.1016/j.comnet.2017.05.004.
- [28] L. Chen, Y. Zhang, Y. Chen, B. Li, and W. Liu, "Negative influence blocking maximization with uncertain sources under the independent cascade model," *Inf Sci (N Y)*, vol. 564, pp. 343–367, Jul. 2021, doi: 10.1016/j.ins.2021.02.063.
- [29] W. Zhu, W. Yang, S. Xuan, D. Man, W. Wang, and X. Du, "Location-Aware Influence Blocking Maximization in Social Networks," *IEEE Access*, vol. 6, pp. 61462–61477, 2018, doi: 10.1109/ACCESS.2018.2876141.
- [30] H. Huang, H. Shen, Z. Meng, H. Chang, and H. He, "Community-based influence maximization for viral marketing," *Applied Intelligence*, vol. 49, no. 6, pp. 2137–2150, Jun. 2019, doi: 10.1007/s10489-018-1387-8.
- [31] L. Li, Y. Liu, Q. Zhou, W. Yang, and J. Yuan, "Targeted influence maximization under a multifactor-based information propagation model," *Inf Sci (N Y)*, vol. 519, pp. 124–140, May 2020, doi: 10.1016/j.ins.2020.01.040.
- [32] T. K. Biswas, A. Abbasi, and R. K. Chakraborty, "Selecting a Cost-Effective Seed for Maximizing the Social Influence Under Real-life Constraints."
- [33] F. Kazemzadeh, A. A. Safaei, and M. Mirzazadee, "Influence maximization in social networks using effective community detection," *Physica A: Statistical Mechanics and its Applications*, vol. 598, Jul. 2022, doi: 10.1016/j.physa.2022.127314.
- [34] A. Uthayasuriyan, G. Hema Chandran, U. Kavvin, S. H. Mahitha, and G. Jeyakumar, "A Comparative Analysis on Influence Maximization Models in Social Networks," in *2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023, Institute of Electrical and Electronics Engineers Inc.*, 2023, pp. 456–461. doi: 10.1109/InCACCT57535.2023.10141772.
- [35] 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022.
- [36] Z. Aghaee and S. Kianian, "Influence maximization algorithm based on reducing search space in the social networks," *SN Appl Sci*, vol. 2, no. 12, Dec. 2020, doi: 10.1007/s42452-020-03812-w.
- [37] E. Zhu, L. Yang, and Y. Xu, "A Local Search Algorithm for the Influence Maximization Problem," *Front Phys*, vol. 9, Oct. 2021, doi: 10.3389/fphy.2021.768093.
- [38] Y. Li, J. Fan, Y. Wang, and K. L. Tan, "Influence Maximization on Social Graphs: A Survey," *IEEE Trans Knowl Data Eng*, vol. 30, no. 10, pp. 1852–1872, Oct. 2018, doi: 10.1109/TKDE.2018.2807843.
- [39] Y. Y. Ko, K. J. Cho, and S. W. Kim, "Efficient and effective influence maximization in social networks: A hybrid-approach," *Inf Sci (N Y)*, vol. 465, pp. 144–161, Oct. 2018, doi: 10.1016/j.ins.2018.07.003.
- [40] S. Kim et al., "Scalable and parallelizable influence maximization with Random Walk Ranking and Rank Merge Pruning," *Inf Sci (N Y)*, vol. 415–416, pp. 171–189, Nov. 2017, doi: 10.1016/j.ins.2017.06.018.
- [41] X. Liu, S. Wu, C. Liu, and Y. Zhang, "Social network node influence maximization method combined with degree discount and local node optimization," *Soc Netw Anal Min*, vol. 11, no. 1, Dec. 2021, doi: 10.1007/s13278-021-00733-3.
- [42] A. Alamsyah and B. Rahardjo, "Social Network Analysis Taxonomy Based on Graph Representation."