IMT: Selection of Top-k Nodes based on the Topology Structure in Social Networks

Hamid Ahmadi Beni
Department of Information Technology and Communications
Azarbaijan Shahid Madani University Tabriz,
Tabriz, Iran
h.ahmadi@azaruniv.ac.ir

Zahra Aghaee*
Faculty of Computer Engineering, Shahid Rajaee Teacher Training University (SRTTU) ,
Tehran, Iran
z.aghayii@sru.ac.ir

Asgarali Bouyer
Department of Information Technology and Communications
Azarbaijan Shahid Madani University Tabriz,
Tabriz, Iran
a.bouyer@azaruniv.ac.ir

Mehdi Vahidipour
Faculty of Electrical and Computer Engineering, Department of Computer, University of Kashan,
Kashan, Iran
vahidipour@kashanu.ac.ir

***Abstract*— Influence maximization is a problem based on diffusion and probability in social networks with the aim of finding the least k node with the most influence. These nodes play an essential role in the diffusion process. However, the influence maximization problem faces two essential challenges of time efficiency and optimal selection of the seed nodes. To solve these challenges, we proposed an algorithm based on the properties of the graph topology structure and centrality, called IMT (Influence Maximization based on the Topology) algorithm. This algorithm selects the seed nodes from the dense part of the graph that can access more nodes in the shortest distance. Finally, experiments showed that the proposed algorithm outperformed the other algorithms in terms of influence spread and running time.**

Keywords—Influence Maximization Problem; Information Diffusion; Influence Spread; Graph Topology; Centrality

I. INTRODUCTION

In recent years, people's interest in using social networks such as Facebook, Twitter, and LinkedIn has increased for a variety of reasons, including easier communication with others, easier access to new information and news, and even ease of product sale. Therefore, the effect of social networks on individuals and economics is undeniable. As such, these networks provide a good platform for advertising. Advertising companies need to find influential people in social networks to sell a large number of their products in a short time. This process is called viral marketing [1], which is modelled on the Influence Maximization Problem (IMP) in social networks [2].

In the IMP, individuals or nodes with a specified number of social networks or graphs are selected with the greatest effect on the network [3]. In other words, the goal of IMP is to diffuse information to a large number of people on a social network in a short time by a set of effective individuals (seed nodes) [4, 5]. The process of diffusion from person to person is similar to a tweet on the Twitter social network when a specific event occurs.

Research in the field of IMP has presented various information diffusion models [2, 6-13], the most commonly used models of which are the Independent Cascade Model (ICM) [2] and Linear Threshold Model (LTM) [2]. The IMP based on these two models is an NP-hard with submodular and monotonic properties [2]. Therefore, the algorithms presented by the researchers solve this problem to some extent. The algorithms have a number of major drawbacks, including (1) In most algorithms proposed for the IMP, the high centrality nodes are not considered while these nodes can be considered as an important and effective indicator in the diffusion process. (2) Most of them have high running time on large-scale social networks. (3) They do not have optimal selection in seed nodes and therefore do not have suitable influence spread. Hence, it is necessary to provide methods for reducing the running time on the on large-scale graphs as the scale of social network increases. Consequently, this article provided a new IMT (Influence Maximization based on the Topology) algorithm under an ICM. The IMT algorithm is based on the properties of the graph topology structure and centrality with respect to the part of the graph that is dense. As a result, application of this feature to the IMT algorithm reduces the running time and achieves high influence spread compared to the topology-based algorithms on all four datasets.

In summary, our main contributions in this article were:

(1) Provision of a new algorithm by examining the degree of neighbour nodes at different levels for the IMP, which increases the algorithm efficiency.

(2) Provision of a new algorithm that selects the seed nodes from the dense part of the graph using a combination of the properties of the graph topology structure and harmonic centrality.

(3) Comprehensive experiments on the real-world datasets of social networks in terms of influence spread and running time, indicating the effectiveness of our algorithm in comparison to other topology-based algorithms.

Section II briefly discusses the related works and Section II presents the proposed method. Section IV analyzes and evaluates the experiment. Finally, Section V makes conclusion and suggests some future works.

II. RELATED WORK

IMP is one of the hard optimization problems that first addressed by Domingos and Richardson [14]. Afterwards, Kempe et al. proved that this was an NP-hard problem [2]. These researchers have formulated this problem under ICM and LTM [2]. Accordingly, the influence spread in the IMP was calculated by (1) [15]. In this Equation and by considering the graph $G(V, E)$ under a diffusion model with the goal of selecting k seed nodes, an average number of nodes activated by $\sigma(S)$ is calculated. S^* denotes the seed nodes with the maximum activation in the graph.

978-1-7281-1051-6/20/$31.00 ©2020 IEEE

$$S^* = argmax_{S \subset V, |S|=k} \sigma(S) \quad (1)$$

Kempe et al. proposed a greedy algorithm to solve the IMP [2]. Although this algorithm offered optimal influence spread, it had a very high running time, especially for large-scale graphs due to the use of Monte Carlo simulation. Therefore, to improve the running time of the greedy algorithm, various algorithms were presented, including greedy algorithms CELF [16], CELF++ [17], NewGreedyIC [4] and NewGreedyWC [4], but these algorithms still failed to handle the running time on large-scale graphs. Moreover, various heuristic algorithms, such as HighDegree [2] and SingleDiscount [4] algorithms, have been developed. These algorithms improved the running time as well as the influence spread. Then, researchers proposed topology-based algorithms, including CI [18], VoteRank [19], and LIR [20] algorithms to handle the running time on the large-scale graphs. The CI algorithm selected the seed nodes using the localization of influence spread. The main disadvantage of this algorithm was the dependence of the influence spread calculations on the localization criteria. The VoteRank algorithm was introduced to improve the CI algorithm. It was also a voting-based algorithm. The main disadvantage of the algorithm was the lack of an optimal approximation guarantee. In addition, LIR algorithm was presented to improve the NewGreedyIC algorithm. This algorithm also had very fast running time but did not guarantee optimal approximation and had low influence spread. In this regard, Qiu et al. proposed the PHG algorithm to improve the greedy algorithm [21]. The algorithm was based on the community detection and offered the influence spread close to the greedy algorithm, but did not guarantee optimal approximation. Moreover, Ahmadi Beni et al. developed the TI-SC algorithm to improve the CI, LIR, and PHG algorithms [22]. This algorithm had efficient performance in high rich-club social networks but with high running time in the large-scale social networks.

III. Proposed Algorithm

In the IMT algorithm, an approach was presented on the basis of the properties of the graph topology structure and centrality in the graph. Since diffusion in social networks depended on the graph structure, it was necessary to consider the intuitive parameter statistics and graph topology structure properties in the IMT algorithm for selecting the influential nodes in the graph and the node distance factor. These attributes helped to select the influential nodes from parts of the graph that were dense. Fig 1. depicts the general framework of the IMT algorithm.

According to Fig 1, in the IMT algorithm, for every node v_i, the equation of ITC_{v_i} was calculated based on the topology structure and centrality. $l1$ and $l2$ in (2) are the number of the nodes at level one and level two of the node v_i. D_j represents the degree of the node v_j, which is in the neighbourhood of the node v_i. LI_{v_i} refers to the local index of each node v_i. This index stands for the number of nodes at level one of node v_i that their degree is more than the degree of node v_i. Finally, H_{v_i} is the harmonic centrality of each node v_i.

$$ITC_{v_i} = \frac{\sum_{j=1}^{l1} D_j \times \sum_{j=1}^{l2} D_j}{LI_{v_i} + 1} + H_{v_i} \quad (2)$$

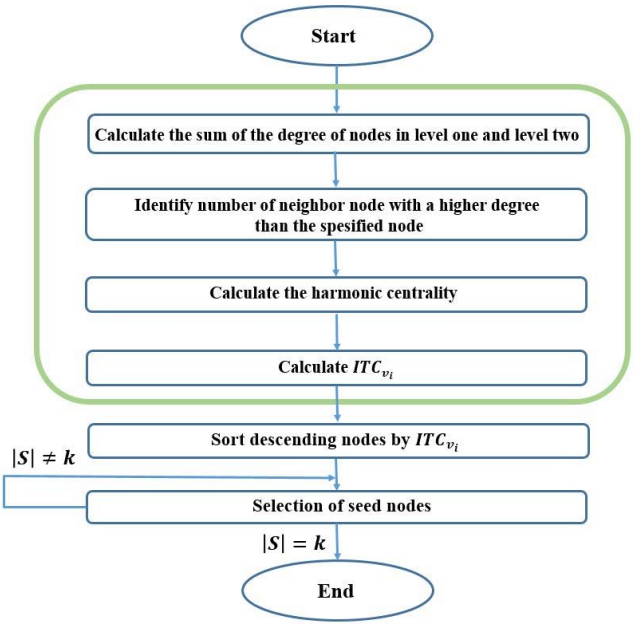

Fig. 1. The general framework of the IMT algorithm

$\sum_{j=1}^{l1} D_j$ is the sum of the degree of nodes at level one of each node v_i. For example, according to the graph in Fig. 2, the sum of the degrees of nodes 9, 11, and 19 were calculated to calculate the sum of the degree of nodes at level one of the node 10.

$\sum_{j=1}^{l2} D_j$ is the sum of the degree of the nodes at level two of each node v_i. For example, according to the graph in Fig 2, the sum of the degree of blue nodes was calculated to calculate the sum of the degree of nodes at level two of node 10.

The harmonic centrality H_{v_i} is the sum of the reciprocal of the shortest path distances from all other nodes of the graph [23]. This centrality selected the nodes reachable to the greater number of nodes in the graph with the least distance and the least number of interface nodes. Equation (3) calculates H_{v_i} for each node v_i [23]. $d(v_i, u_j)$ is the shortest distance between the nodes v_i and u_j.

$$H_{v_i} = \sum_{v_i \neq u} \frac{1}{d(v_i, u)} \quad (3)$$

Following the calculation of ITC_{v_i} equation for all graph nodes, the nodes were sorted in descending order by the ITC_{v_i}. The k nodes were selected from the top of the sorted list as the seed nodes. Then, the influence spread was calculated under the ICM. Algorithm 1 shows the pseudo-code of the IMT algorithm.

According to lines 2-4 of Algorithm 1, ITC_{v_i} equation was calculated for each node v_i. In line 5, the nodes were sorted in the descending order according to the ITC_{v_i}, and in line 6, k nodes were selected as the seed nodes from the top of the sorted list. Moreover, in lines 7-10 for every node v_m in the set N, the influence spread under the ICM was calculated (σ is the influence spread under the ICM), and finally, the set S was returned as the set of the seed nodes.

The time complexity in the IMT algorithm consisted of two steps. In step 1, ITC_{v_i} equation was calculated for all

graph nodes and then the k seed node was selected with the maximum value of ITC_{v_i}, which had the time complexity $O(m)$. m is the number of the edges in graph G. In step 2, the influence spread of the seed nodes was calculated under an ICM with the time complexity of $O(kR)$. k is the number of seed nodes and R is the number of Monte Carlo simulations. In general, the time complexity of the IMT algorithm is $O(m + kR)$.

IV. Experiments

In this section, four datasets were used to evaluate the effectiveness and efficiency of the IMT algorithm compared with the other three well-known algorithms. Simulation was programmed in Python language and implemented on a computer with 2.5GHz Intel Core i5 CPU-3230M and 16 GB memory. TABLE I reports the specifications of four real social network datasets, including Email [24], Route views [25], Sister cities [25], and As-22july06 [25].

According to the research design, performance of the IMT algorithm was compared with three based-topology algorithms of LIR, CI, and VoteRank, which have been successfully applied to social networks. According to [20], in the LIR algorithm, the nodes with $0 - LI$ were selected for the Route views, Sister cities, and As-22july06 datasets, while in the Email dataset, the nodes with $0 - LI$ were low, then the nodes with $1 - LI$ were selected. According to [18], localization parameter $l = 3$ was used in the CI algorithm. All algorithms were implemented with the parameter Monte Carlo simulation $R = 1000$ and the activation probability $p = 0.01$.

A. Diffusion model

Researchers have provided various types of diffusion models such as Independent Cascade Model (ICM)[2], Linear Threshold Model (LTM) [2], Weighted Cascade Model (WCM) [4], Susceptible Infected Recovered (SIR) [26], Susceptible Infected Susceptible (SIS) [27], and so forth to model the information diffusion in social networks. However, one of the most commonly used models is ICM, which is the IMT algorithm under this model. In this model, nodes have two active or inactive states. Node v can activate node u with probability $p(v,u)$. Moreover, activation is done only once, if node v cannot activate node u; thus, node v has no chance to activate node u. Therefore, the activation process continues until another new node cannot be activated.

B. Evaluation metrics

Two metrics were used to evaluate the efficiency of the IMT algorithm: influence spread and running time. The influence spread means measuring the accuracy of the seed nodes in the information diffusion. This is the average number of the activated nodes for each iteration of the Monte Carlo simulation. The running time means the time elapsed to find the seed nodes.

C. Influence spread results

To further investigate the effectiveness and efficiency of the IMT algorithm, the influence spread was evaluated in real social networks where the x-axis represents the number of the seed nodes, while the y-axis represented the overall influence spread. The results in four datasets reflected that the IMT algorithm completely outperformed other compared algorithms in terms of influence spread. The LIR algorithm

Algorithm 1: IMT algorithm

Input: $G = (V, E), V = \{v_1, v_2, v_3, \dots, v_n\}$
Output: S
1: initialize $S \longrightarrow \emptyset$;
2: **for each** v_i in V **do**
3: $ITC_{v_i} = \frac{\sum_{j=1}^{l1} D_j \times \sum_{j=1}^{l2} D_j}{LI_{v_i}+1} + H_{v_i}$
4: **end for**
5: sort the values of ITC_{v_i} in descending order
6: select the number of k nodes with the largest ITC_{v_i} and add to set N
7: **for each** v_m in N **do**
8: $S \leftarrow v_m$
9: $\sigma \leftarrow \sigma(v_m \cup S)$
10: **end for**
11: **Return** S *and* σ *//S is seed set nodes.*

TABLE I. Specifications of four datasets

Dataset	Email	Route views	Sister cities	As-22july06
# Node	1k	6k	14k	23k
# Edge	5k	13k	20k	48k
Max Degree	209	1459	99	2390
Min Degree	1	1	1	1

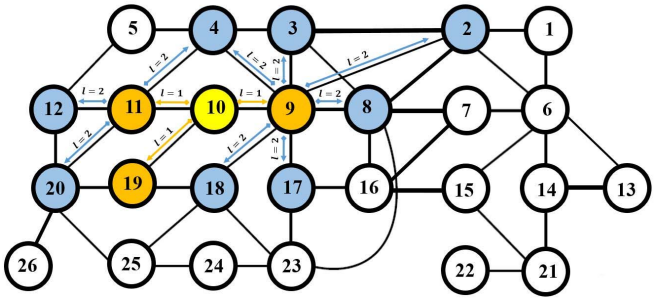

Fig. 2. An example of the calculation of the number of nodes at level one and level two for the IMT algorithm

showed the worst performance on all networks. Moreover, IMT algorithm performed well on all datasets. Based on Fig. 3, IMT and VoteRank algorithms exhibited the same influence spread. In addition, significant gaps of influence spread could be observed between the IMT algorithm and other algorithms in Fig. 4. According to Fig. 5, the influence spread of CI, VoteRank, and LIR algorithms was lower than the IMT algorithm. For example, on the Sister cities dataset with $k = 20$, the IMT algorithm achieved the influence spread of 31.5639, while the influence spread of CI, VoteRank, and LIR were 28.9749, 29.2649, and 26.2129, respectively. As shown in Fig 6, the IMT algorithm exhibited the same influence spread. From the results on the real-world datasets, it could generally concluded that the IMT algorithm had better efficiency than the state-of-the-art algorithms in finding the influential seed nodes. Finally, the VoteRank algorithm enjoyed the second-best performance.

D. Running time results

Fig. 7 shows the running time of different algorithms on the four datasets. As seen in Fig. 7, the running time is the time of selection of $k = 30$ seed nodes. Based on the findings, the running time of LIR and VoteRank were low on all datasets but could not provide any performance guarantee in terms of influence spread. However, the worst running time was for the CI algorithm and the IMT algorithm provided high influence spread along with acceptable running time compared to other algorithms.

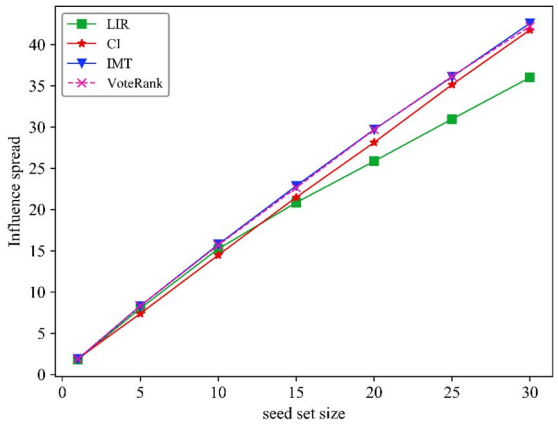

Fig. 3. Influence spread of different algorithms on the Email dataset

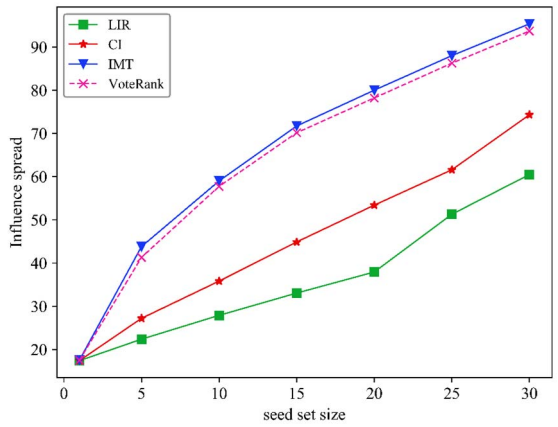

Fig. 4. Influence spread of different algorithms on the Route views dataset

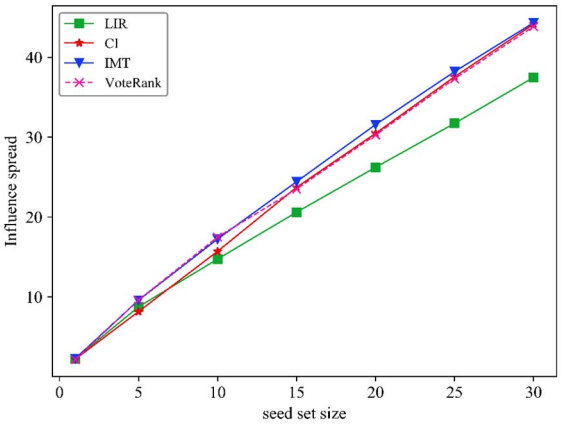

Fig. 5. Influence spread of different algorithms on the Sister cities dataset

V. Conclusion and Future Work

This paper proposed the IMT algorithm to select the set of seed nodes for IMP. The IMT algorithm used a framework based on the properties of the graph topology structure and harmonic centrality to improve performance. According to the algorithm process, the seed nodes were selected from the dense part of the graph. Moreover, experimental results in the four datasets showed the effectiveness and efficiency of the proposed algorithm under the ICM. Therefore, future studies should investigate the IMT algorithm on LTM. In addition, researchers should study the influence of other topology properties such as similarity criterion and clustering coefficient on the selection of seed nodes.

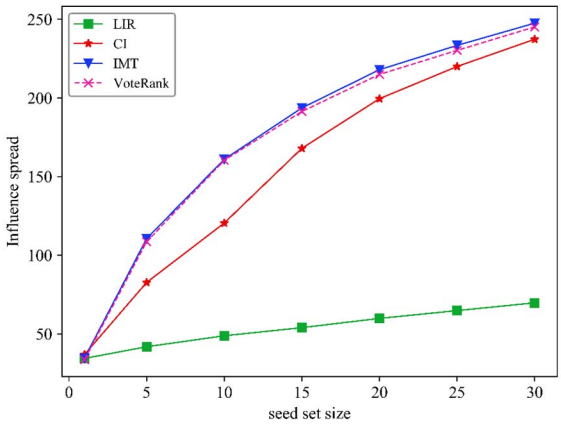

Fig. 6. Influence spread of different algorithms on the As-22july06 dataset

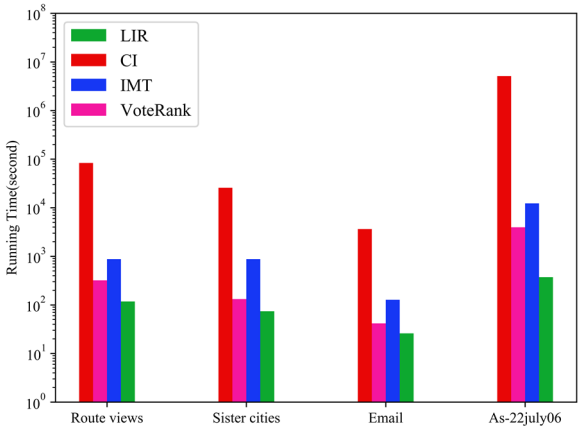

Fig. 7. Running time of different algorithms on four real-world datasets

References

[1] W. Yang, L. Brenner, and A. Giua, "Influence Maximization in Independent Cascade Networks Based on Activation Probability Computation," *IEEE Access,* vol. 7, pp. 13745-13757, 2019.

[2] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003: ACM, pp. 137-146.

[3] A. Talukder, M. G. R. Alam, N. H. Tran, D. Niyato, and C. S. Hong, "Knapsack-based reverse influence maximization for target marketing in social networks," *IEEE Access,* vol. 7, pp. 44182-44198, 2019.

[4] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009: ACM, pp. 199-208.

[5] S. S. Singh, A. Kumar, K. Singh, and B. Biswas, "C2IM: Community based context-aware influence maximization in social networks," *Physica A: Statistical Mechanics and its Applications,* vol. 514, pp. 796-818, 2019.

[6] A. Goyal, W. Lu, and L. V. Lakshmanan, "Simpath: An efficient algorithm for influence maximization under the linear threshold model," in *2011 IEEE 11th international conference on data mining*, 2011: IEEE, pp. 211-220.

[7] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: a survey," *Journal of Network and Computer Applications,* vol. 106, pp. 17-32, 2018.

[8] W. Wang and W. N. Street, "Modeling and maximizing influence diffusion in social networks for viral marketing," *Applied network science,* vol. 3, no. 1, p. 6, 2018.

[9] A. Bozorgi, S. Samet, J. Kwisthout, and T. Wareham, "Community-based influence maximization in social networks under a competitive linear threshold model," *Knowledge-Based Systems,* vol. 134, pp. 149-158, 2017.

[10] Y.-S. Kwon, S.-W. Kim, S. Park, S.-H. Lim, and J. B. Lee, "The information diffusion model in the blog world," in *Proceedings of the 3rd workshop on social network mining and analysis*, 2009: ACM, p. 4.

[11] H. Wang, F. Wang, and K. Xu, "Modeling information diffusion in online social networks with partial differential equations," *arXiv preprint arXiv:1310.0505,* 2013.

[12] J.-J. Cheng, Y. Liu, B. Shen, and W.-G. Yuan, "An epidemic model of rumor diffusion in online social networks," *The European Physical Journal B,* vol. 86, no. 1, p. 29, 2013.

[13] Y. Lin, J. C. Lui, K. Jung, and S. Lim, "Modelling multi-state diffusion process in complex networks: theory and applications," *Journal of Complex Networks,* vol. 2, no. 4, pp. 431-459, 2014.

[14] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001: ACM, pp. 57-66.

[15] W. Chen, L. V. Lakshmanan, and C. Castillo, "Information and influence propagation in social networks," *Synthesis Lectures on Data Management,* vol. 5, no. 4, pp. 1-177, 2013.

[16] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007: ACM, pp. 420-429.

[17] A. Goyal, W. Lu, and L. V. Lakshmanan, "Celf++: optimizing the greedy algorithm for influence maximization in social networks," in *Proceedings of the 20th international conference companion on World wide web*, 2011: ACM, pp. 47-48.

[18] F. Morone, B. Min, L. Bo, R. Mari, and H. A. Makse, "Collective influence algorithm to find influencers via optimal percolation in massively large social media," *Scientific reports,* vol. 6, p. 30062, 2016.

[19] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, "Identifying a set of influential spreaders in complex networks," *Scientific reports,* vol. 6, p. 27823, 2016.

[20] D. Liu, Y. Jing, J. Zhao, W. Wang, and G. Song, "A fast and efficient algorithm for mining top-k nodes in complex networks," *Scientific reports,* vol. 7, p. 43330, 2017.

[21] L. Qiu, W. Jia, J. Yu, X. Fan, and W. Gao, "PHG: A Three-Phase Algorithm for Influence Maximization Based on Community Structure," *IEEE Access,* vol. 7, pp. 62511-62522, 2019.

[22] H. A. Beni and A. Bouyer, "TI-SC: top-k influential nodes selection based on community detection and scoring criteria in social networks," *Journal of Ambient Intelligence and Humanized Computing,* pp. 1-20.

[23] P. Boldi and S. Vigna, "Axioms for centrality," *Internet Mathematics,* vol. 10, no. 3-4, pp. 222-262, 2014.

[24] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical review E,* vol. 68, no. 6, p. 065103, 2003.

[25] J. Kunegis, "Konect: the koblenz network collection," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013: ACM, pp. 1343-1350.

[26] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, "The modeling of global epidemics: Stochastic dynamics and predictability," *Bulletin of mathematical biology,* vol. 68, no. 8, pp. 1893-1921, 2006.

[27] K. Saito, M. Kimura, and H. Motoda, "Discovering influential nodes for SIS models in social networks," in *International Conference on Discovery Science*, 2009: Springer, pp. 302-316.