IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 18, 2017, accepted October 16, 2017, date of publication October 20, 2017, date of current version November 7, 2017.

*Digital Object Identifier 10.1109/ACCESS.2017.2764750*

# A Novel Centrality Cascading Based Edge Parameter Evaluation Method for Robust Influence Maximization

**XIAOLONG DENG[1], YINGTONG DOU[2], TIEJUN LV[3], (Senior Member, IEEE), AND QUOC VIET HUNG NGUYEN [4]**

[1]Key Laboratory of Trustworthy Distributed Computing and Service of Education Ministry, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]International School, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3]Information and Communication Engineering School, Beijing University of Posts and Telecommunications, Beijing 100876, China
[4]School of Information and Communication Technology, Griffith University, Brisbane, QLD 4215, Australia

Corresponding author: Xiaolong Deng (shannondeng@bupt.edu.cn)

This work was supported in part by the Philosophy and Social Science Project of Education Ministry under Grant 15JZD027, in part by the National Key Research and Development Program of China under Grant 2016YFC0800808, in part by the National Culture Support Foundation Project of China under Grant 2013BAH43F01, and in part by the National 973 Program Foundation Project of China in social network analysis under Grant 2013CB329605.

**ABSTRACT** The research of social influence is an important topic in online social network analysis. Influence maximization is the problem of finding $k$ nodes that maximize the influence spread in a specific social network. Robust influence maximization is a novel topic that focuses on the uncertainty factors among the influence propagation models and algorithms. It aims to find a seed set with a definite size that has robust performance with different influence functions under various uncertainty factors. In this paper, we propose a centrality-based edge activation probability evaluation method in the independent cascade model. We consider four different types of centrality measurement methods and add a modification coefficient to evaluate the edge probability. We also propose two algorithms, called NewDiscount and GreedyCIC, by incorporating the edge probability space into previous algorithms. With extensive experiments on various real online social network data sets, we find that our PageRank-based greedy algorithm has the best influence spreads and lowest running times, compared with other algorithms, on some large data sets. The experiment for evaluating the robustness performance shows that all algorithms have optimal robustness performance when the modification coefficient is set to 0.01 under the independent cascade model. This result suggests some further research directions under this model.

**INDEX TERMS** Social networks, influence maximization, robust optimization, information diffusion.

## I. INTRODUCTION

Online social network analysis is developing with the promotion of online social web services. This analysis focuses on information diffusion, user behaviour and other special features of the online social network. With the development of computational technology, increasingly novel algorithms are applied in social network analysis. Meanwhile, some new topics, such as viral marketing, privacy protection, social recommendation and fake news detection, are generated.

Influence Maximization is a specific branch of social network analysis that is derived from utilizing the Word of Mouth (WoM) effect in online viral marketing. Kempe *et al.* first transform this problem into mathematical models called the *Independent Cascade Model* and the *Linear Threshold Model* [1]. These models aim to find a seed set with size $k$ that maximizes the influence propagation in a specific network. To solve this problem, Kempe *et al.* design a greedy method and a heuristic method, which are two basic algorithmic frameworks for Influence Maximization.

Based on the work of Kempe *et al.*, many studies have been carried out to optimize the process of finding the seed set. Most of these works are based on the *Independent Cascade Model* [2]–[6]. These papers focus on how to improve the efficiency of algorithms by taking less time to find the most influential nodes in the network. Some works add novel machine learning algorithms and other social network

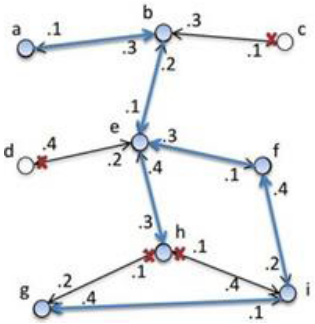



**FIGURE 1.** The independent cascade model [21].

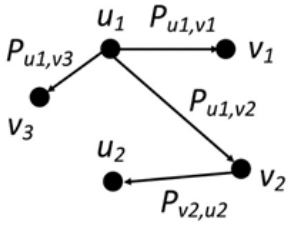



**FIGURE 2.** Influence diffusion process of the CIC model.

analysis techniques to the *Influence Maximization problem* [7], [8]. These works all set the edge activation probability as the ground truth, i.e., 0.01. To evaluate the true edge activation probability, many studies, such as [9]–[13], propose various methods. Since there are many uncertainty factors in influence propagation models, it is still hard to obtain the correct edge probability.

Together with the uncertainty in the edge probability, other uncertainty factors in the *Influence Maximization problem*, such as the variety of models and algorithms, are considered by researchers. He and Xinran [14] and He and Kempe [15] define the uncertainty factors of *Influence Maximization problems* as noise. They propose a new topic called the Robust Influence Maximization (RIM) problem. The goal is to find the algorithm that has the most robust performance in the *Influence Maximization problem* with a robust optimization objective. With the optimization objective of He *et al.*, Chen *et al.* focus on the edge probability and use different sampling methods to evaluate the edge probability [16]. Lowalekar *et al.* propose a new robust optimization objective and evaluate various algorithms under the objective [17].

In this article, we propose a new centrality-based edge activation probability evaluation method under the *Independent Cascade Model*. In the *Independent Cascade Model*, one node influences another node according to the edge probability $p$, which is set to a default value. In the Centrality-based Independent Cascade (CIC) model, a node with higher centrality has higher probability to influence a node with lower centrality. The edge activation probability (also called the edge parameter) depends on the centralities of the nodes on both sides of the edge. To simulate the real situation, we add a modification coefficient $\delta$ to simulate the noise. Meanwhile, we investigate the general centrality measurement methods and select four different centrality measurement methods (Degree, PageRank, Eccentric and Closeness) as the noise in the model. For the RIM problem, we add the edge parameter space to the robust optimization objective proposed by He and Kempe [15]. We improve the NewGreedyIC and DegreeHeuristic algorithms by incorporating the edge probability space into the original algorithm. The two new algorithms that we propose are called GreedyCIC and NewDiscount.

In the experiment, we take four datasets from different online social networks with different sizes and statistical features that could represent the general noise in the problem. We first investigate the influence spread of NewDiscount under four datasets with different sizes of seed set. The parameter space of the algorithms is calculated by different centrality measurement methods. Then, we use visualization tools to show the seed sets selected by different methods. We also investigate the influence spreads of different algorithms and their running times under different noise. Lastly, we study the robustness performance of algorithms.

In this paper, Section II describes the current related work. Section III introduces the modelling procedure and problem definition. Section IV presents the details of our algorithm. Section V gives the experimental results, and Section VI presents the conclusion and discusses future work.

## II. RELATED WORKS

Kempe *et al.* first addressed the *Influence Maximization problem* using specific mathematical models [1]. It is the problem of finding a set of $k$ nodes as influence propagation initialization nodes in a specific social network graph that maximizes the number of finally influenced nodes. The social graph can be defined as $G(V, E)$, where $V$ denotes the vertices (users) of the graph and $E$ denotes the edges (connections) among vertices. The *influence maximization problem* is defined as:

$$S_\theta^* = \frac{\text{argmax}}{S \subseteq V, |S| = k} \sigma_\theta(S), \tag{1}$$

where $\sigma_\theta(S)$ is the influence propagation function, and the objective is to find the seed set $S$ that maximizes the final quantity of influenced nodes.

Kempe *et al.* proposed two basic influence diffusion models in the online social network [1]: the *Independent Cascade Model* and *Linear Threshold Model* They also prove that the *Influence Maximization problem* is an NP-hard problem. To solve this problem, there are two basic algorithm models. The first model is the greedy algorithm. It traverses all the nodes and adds the one that maximizes the influence of the seed set; then, it traverses other nodes until it finds $k$ nodes to compose the seed set. This method has the disadvantage that the efficiency is very low. Another algorithm model is the heuristic algorithm. It selects a random set of nodes in the social graph as the seed set. Then, it calculates the final influence effect to find the best seed set. It has lower

**TABLE 1. Major centrality measurement methods.**

| Name | Type | Description |
|---|---|---|
| Degree Centrality | Local Measurement | Degree is defined as the number of links incident upon a node |
| Closeness Centrality | Global Measurement | Closeness is the average length of the shortest path between the node and all other nodes in the graph |
| Betweenness Centrality | Global Measurement | Betweenness means the number of times a node is passed by the shortest paths in a whole graph |
| Eigenvector Centrality | Global Measurement | Eigenvector centrality assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes |
| Katz Centrality | Local Measurement | Katz centrality measures the total number of nodes that can be connected through a path, while the contributions of distant nodes are penalized |
| PageRank Centrality | Global Measurement | PageRank Centrality is a little different from Eigenvector Centrality and is generalized from the PageRank method in web information retrieval |

accuracy but is faster than the greedy algorithm. In 2003 [1], Kempe *et al.* proposed the Basic Greedy and heuristic algorithms based on centrality and degree.

In 2007, Leskovec *et al.* [6] proposed a CELF method that utilizes the sub-modularity of the influence objective function. This method reduces the complexity of the greedy algorithm and improves its efficiency. In 2009, Chen *et al.* [2] proposed a Degree Discount algorithm that outperformed the traditional heuristic algorithms by an order of magnitude in speed. It brings the *influence maximization problem* to a new stage. In 2010, Chen *et al.* focused on the scalability of influence maximization algorithms and proposed an efficient heuristic scalable algorithm [3]. In 2011, Goyal *et al.* optimized the CELF algorithm and proposed a CELF++ algorithm [4]. Some papers present work on the *linear threshold model* and make remarkable contributions [3], [18], [19]. In 2013, Barbieri *et al.* built a new influence propagation model based on the topic detection method [7]. Wang et al. applied the community detection method to the *influence maximization problem* [8].

Since the previous research about social influence maximization focuses on many models and influence functions, the experiments are based on different datasets. There are many unstable factors in the influence diffusion models, especially those dealing with the social *influence maximization problem*. Chen et al. first discussed this issue in 2014 [20]. They investigated the instability of the *influence maximization problem* in perturbation models and datasets and proposed an efficient algorithm called the Random Greedy Algorithm that can improve the stability of *influence maximization problem*s to some degree.

Further works have been done by He and Kempe to address the stability of the *influence maximization problem*. They

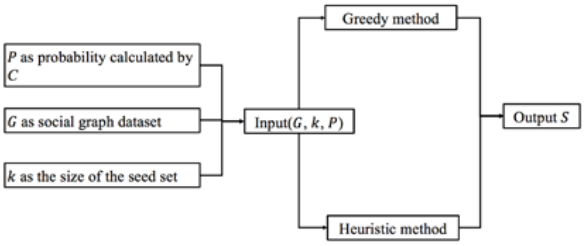



**FIGURE 3. Flow chart of algorithms.**

**TABLE 2. Statistical attributes of datasets.**

| Dataset | Nodes | Edges | Density | Average Degree | Average Clustering Coefficient |
|---|---|---|---|---|---|
| Retweet | 96 | 117 | 0.0257 | 2 | 0.0608 |
| FBMIT | 6.4k | 251.2k | 0.0123 | 78 | 0.2724 |
| Epinions | 26.6k | 100.1k | 0.003 | 7 | 0.1352 |
| Douban | 154.9k | 327.2k | 2.73 | 4 | 0.0161 |

defined a new problem called Robust Influence Maximization [15] in 2015. Robust means the system must adapt the changes in the environment and can work well in multiple situations [14]. In the *Influence Maximization problem*, He and Kempe first classified the noise in social computing into five types in 2015: A. the definition of social ties; B. various mathematical models; C. human behaviour influenced by the environment variables; D. incomplete datasets; and E. uncertainty of parameters [15]. All these types of noises existed in previous research on influence maximization. In detail, the influence function $\sigma$ varies in different influence diffusion models, such as ICM. Furthermore, missing observations also lead to the uncertainty in the values of the parameters of $\sigma$.

To reduce the noise in the *influence maximization problem*, i.e., improve the robustness of the influence functions, He and Kempe propose a robust influence maximization optimization objective:

$$\rho(S) = \min_{\sigma \in S} \frac{\sigma(S)}{\sigma\left(S_{\sigma}^{*}\right)}, \tag{2}$$

where $\sigma$ represents the influence propagation function, $S^*$ is the seed set that maximizes the influence propagation of $\sigma\left(S_{\sigma}^{*}\right)$, and $\rho(S)$ is the optimization objective [15]. Based on the robust influence maximization objective, He *et al.* mainly focus on the perturbation interval discrete-time *independent cascade model* (proposed in their 2014 paper). They prove a very important lemma under this model.

*Lemma 1:* Under the Perturbation Interval model for the Discrete Independent Cascade model, the worst case for the ratio in $\rho$ for any seed set $S_0$ is achieved by making each $P_e$ equal to $l_e$ or $r_e$.

This lemma gives the theoretical basis for calculating the influence propagation under the perturbation interval model. He *et al.* then design three new algorithms: Saturate Greedy, Single Greedy and All Greedy; the last two algorithms are heuristic algorithms. They optimize their algorithms with

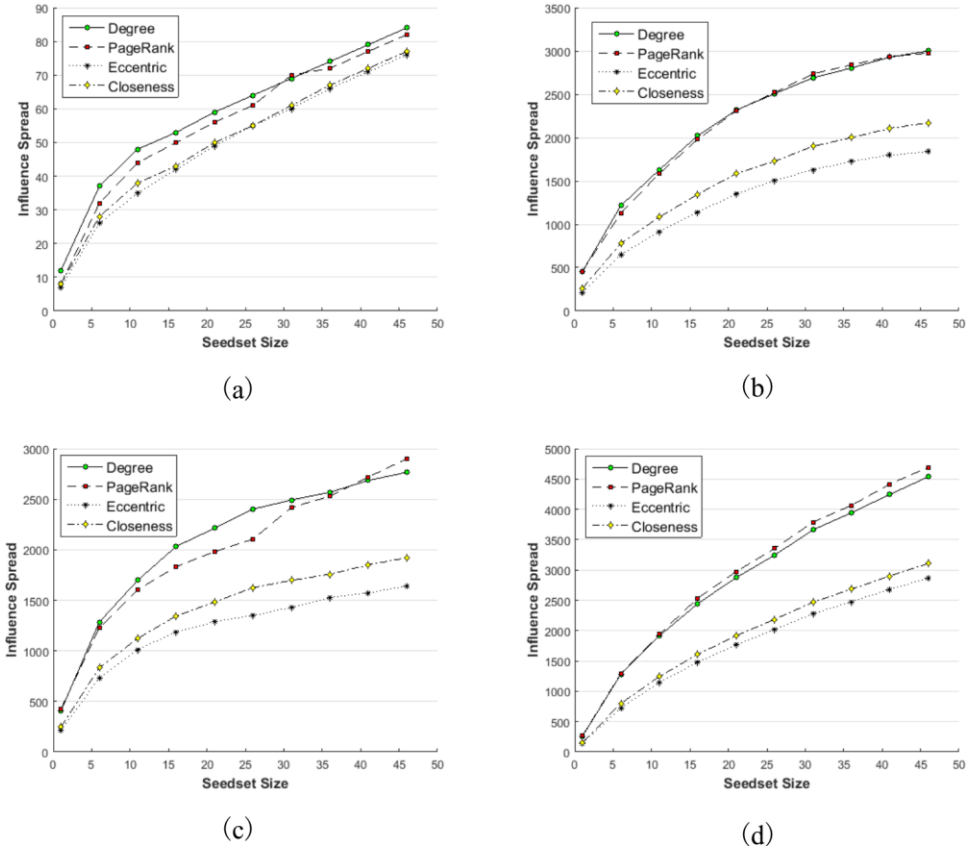



**FIGURE 4.** Influence Spreads for Different Seed Set Sizes (1, 6, 11, 16, 21, 26, 31, 36, 41, 46) Selected by NewDiscount with Different Probability Spaces under Datasets (a) Retweet (96), (b) FBMIT (6.4k), (c) Epinions (26.6k), and (d) Douban (154.9k). The probability space is generated by Degree Centrality, PageRank Centrality, Eccentric Centrality and Closeness Centrality with Formula 7, setting $\lambda = 1$.

the CELF method and ConTinEst methods proposed by Du *et al.* [9]. Then, they perform experiments using several datasets that contain some aspects of the noise and compare the performances of the three algorithms under various experiment setups. They compare the robust and non-robust results by visualization graphs and evaluate the scalability of the algorithms.

Chen *et al.* also discuss the robustness of the *influence maximization problem* in [16]. In their paper, they continue the research under the robust influence maximization objective that was proposed by He *et al.*, which is given as Formula 2, Chen *et al.* carry out some works on measuring the propagation probability in different models. They acknowledge the assumption that the edge activation probability is a distribution between (0, 1). Then, they use different sampling methods to sample the probability during influence diffusion. With the uniform sampling and adaptive sampling methods, they perform experiments with a new greedy algorithm called the Lower-Upper Greedy Algorithm on several datasets. The results show that the parameter uncertainty could substantially influence the performance of the *influence maximization problem*, and adaptive sampling could efficiently improve the robustness of the *influence maximization problem*.

In 2016, in work on robust influence maximization, Lowalekar *et al.* used a different approach from the two works mentioned above [17]. They defined a brand new robust optimization objective:

$$\delta(S, \mathbf{p}) = \max_{S' \in S} -\sigma_P(S), \tag{3}$$

$$\delta^{MR}(S, P) = \max_{p \in P} \delta(S, \mathbf{p}), \tag{4}$$

$$\delta^{MMR}(P) = \max_{S \in \mathbb{S}} \delta^{MR}(S, P). \tag{5}$$

He and Kempe et al. take the Adversarial Noise Model as the basic model [15]. This model assumes that the parameters of the influence function are randomly distributed in a probability space, and the final parameter is decided by the competition results. Lowalekar *et al.* obtain the same Lemma 1 as He et al. The worst influence performance is attained when the parameter is at the boundary of the parameter space, i.e., the regret is the largest at this time. Instead of Formula 2, Lowalekar *et al.* take the minimized maximum regret as the robust optimization objective. The regret is shown in Formula 3, and Formula 4 is the maximum regret. Formula 5 is the minimized maximum regret.

It assumes that the real influence parameters lie in the section $[\check{P}_{u,v}, \hat{P}_{u,v}]$, which represents all the possible seed sets with M nodes. $P = \{[\check{P}_{u,v}, \hat{P}_{u,v}]\}_{e \in E}$ represents the noise of the influence parameters. $S$ and $S'$ are the seed sets with M nodes. $\sigma_P(S)$ is the expected value of influence of parameter $p$.

In addition, Lowalekar *et al.* propose a new evaluation method of influence maximization, i.e., calculating the percentage gap. Lowalekar *et al.* compare some major greedy algorithms experimentally and validate the efficiency of their new method in solving the RIM problem. They conclude that the greedy algorithms perform with good robustness in small-scale real-world social network datasets, even though they do not consider the noise in influence models.

From the state-of-the-art review, it is clear that the robust *influence maximization problem* is a novel topic that has been generated in recent years. The state-of-the-art approach addresses this problem from different viewpoints, but mainly focuses on the noise during the edge activation process, i.e., the uncertainty in the parameters of the influence function. In this paper, we continue the work of these papers and try to investigate the uncertainty in the edge parameters in a novel way and have proposed some new methods.

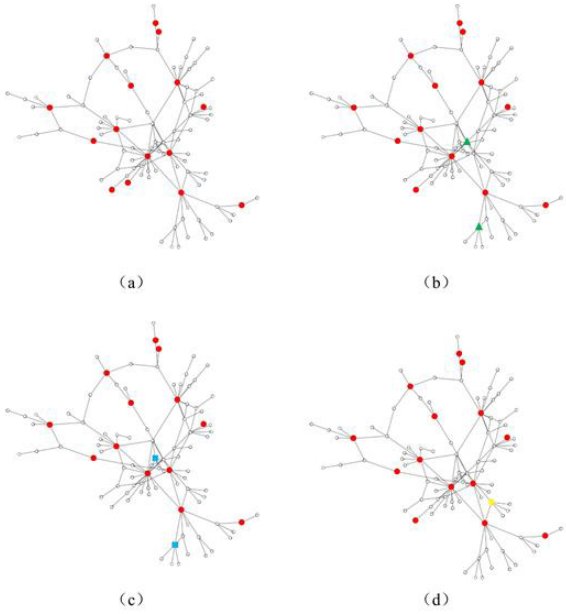



**FIGURE 5.** Fifteen Seed Nodes Selected by Different Centrality Measurement Methods in Retweet. (a) Degree Centrality, (b) PageRank Centrality, (c) Eccentric Centrality, (d) Closeness Centrality. The red circles represent the seed nodes selected by four methods, while the green triangles, blue squares and yellow rhombi represent seed nodes selected by PageRank Centrality, Eccentric Centrality and Closeness Centrality, respectively.

## III. MODELLING AND PROBLEM DEFINITION

### A. CENTRALITY-BASED INDEPENDENT CASCADE (CIC) MODEL

He and Kempe show that there are many noises during the influence diffusion process [15]. Noises are the uncertainty factors of influence diffusion models, including the parameters of the model. To specify the noises, in the *Independent Cascade Model* proposed by Kempe, they use a probability parameter to decide whether the node is going to be activated by another node [1]. In other words, the model randomly generates a variable between (0, 1), and if the variable is larger than the edge activation parameter, then the node at the other side of the edge will be activated, i.e., the influence propagates from one node to another node; otherwise, the influence will not diffuse through this edge. Figure 1 shows the influence propagation process of the IC model. Since the parameter of each edge is difficult to evaluate, most of the algorithms set all the parameters to 0.01 when the influence begins to propagate. Some other models use specific graphs in which the edge parameter is already specified to set up the experiments. Since there are various factors influencing the edge parameter, both of the above methods are of limited utility for simulating real situations. These models have bad robustness to noise in the edge parameter.

To eliminate the noise in the edge parameter, we need to simulate the real influence propagation probability. Some papers extract the influence propagation probability from real-world data [9]–[12]. These probability values always have some deviation comparing with the real probabilities. The evaluation of the edge parameter is still a challenging problem that adds uncertainty to influence propagation models. He and Kempe propose a Perturbation Interval Model in [15]. This model declares that the edge parameter is perturbed in the interval (0, 1). Thus, the probability is not a fixed number such as 0.01, but a random value that is determined at the time we observe it. They also prove that this model attains its worst performance when the parameter is equal to the right or left bound of the interval (Lemma 1). Chen *et al.* consider this model and propose a different sampling method to evaluate the edge parameter in the interval [16].

In this paper, we will also focus on the edge parameter and calculate the parameter by evaluating the centralities of the nodes on the both sides of the edge. This idea is generated from the influence propagation in the real online social network. Taking Sina Micro Blog, which is the largest microblogging web service in China, as an example, people tend to repost messages posted by famous people, i.e., the opinion leadership or KOLs (Key Opinion Leaders) [11]. In Sina Micro Blog, leading members, which are the 'Big V' users, are the KOLs. In social network analysis, there are many ways to identify the KOL in a specific social network. One method considers the connections of a single node in a social graph; we call this the centrality. Centrality in the social network was first proposed by Freeman *et al.* in 1979. It indicates that how important a person (node) is in a social network [12]. Many algorithms have been developed to evaluate the centrality of the graph.

In our model, we calculate the edge parameter with the centralities of the nodes on the two endpoints of the edge.

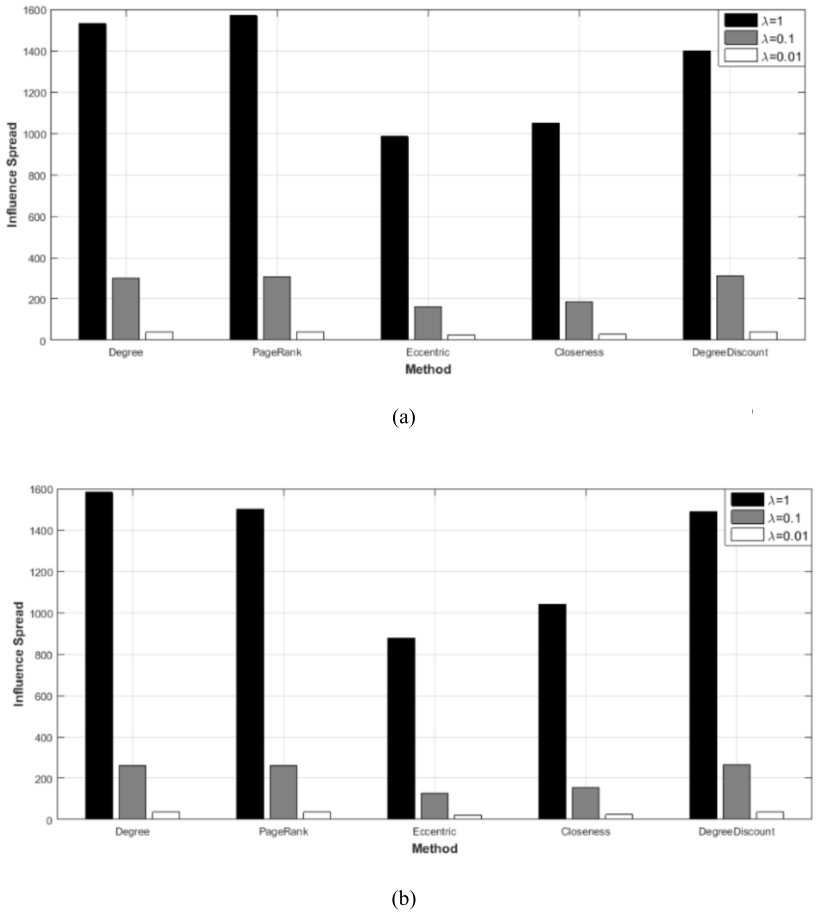



**FIGURE 6.** The Influence Spreads of Seed Nodes Selected by Different Centrality Measurement Methods in NewDiscount with Different λ Values under FBMIT and Epinions. For the DegreeDiscount algorithm, all the edge parameters are equal to λ. (a) Seedset size = 10, with NewDiscount, under FBMIT. (b) Seedset size = 10, with NewDiscount, under Epinions.

Since the influence tends to pass from nodes with high centrality to nodes with low centrality and the edge parameter is between 0 and 1, we use a fraction to represent the centrality difference between nodes:

$$P_{u,v} = \frac{C_u}{C_u + C_v}, \quad for\ C_u + C_v \neq 0. \tag{6}$$

$P_{u,v}$ represents the edge parameter (activation probability) of the edge between $u$ and $v$, and $C_u$ is the centrality of node $u$. We could easily prove that $P_{u,v} > P_{v,u}$ if $C_u > C_v$ and $P$ is always between 0 and 1, which could simulate the edge parameter. In the special situation that the centralities of both nodes are equal to zero, we assume that the edge parameter of the edge between them is zero. This means that influence cannot be passed through this edge. The influence diffusion process is illustrated in Figure 2.

The previous edge parameter in the IC model is a fixed value varying from 0 to 1. Taking the real situation of influence diffusion in online social networks into consideration, the probability of one user reposting another user's message should not be large, e.g., should be less than approximately 0.5. To make the edge parameter calculated by centrality closer to the real value collected from the online social network, we add a modification coefficient λ to Formula 6 to adjust the edge parameter in Formula 7:

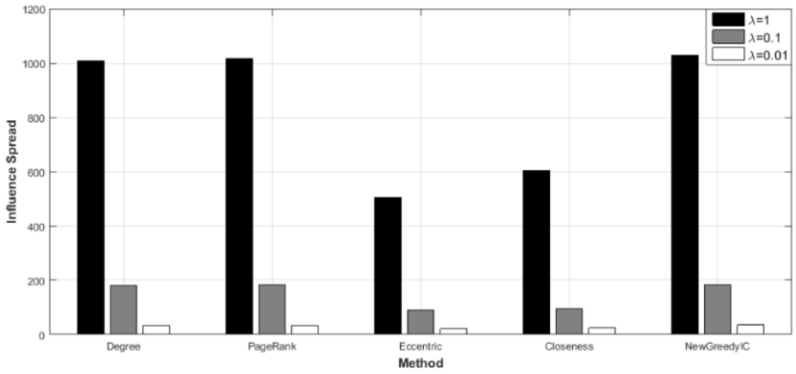
$$P'_{u,v} = \lambda \frac{C_u}{C_u + C_v}, \quad for\ C_u + C_v \neq 0. \tag{7}$$

The modification coefficient λ is set to one of the values in the set (1, 0.1, 0.01). The modified edge parameter can simulate the real probability for different orders of the degree. In the experiment, we will discuss the results of the same algorithm for different values of λ.

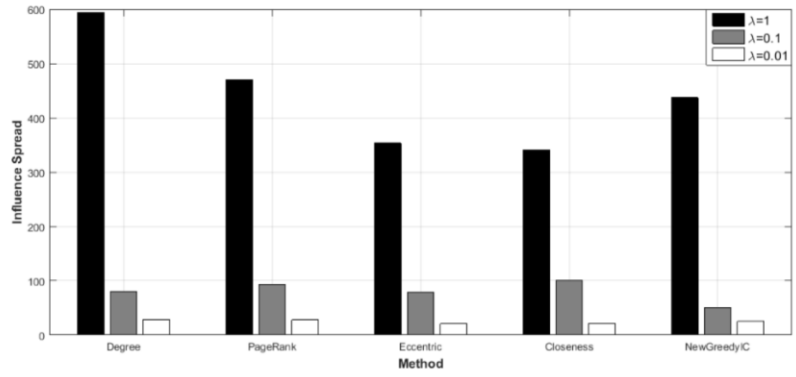
For the directed network, the influence only propagates from one node to the nodes to which it points. Therefore, for the edge from node $u$ pointing to node $v$, the edge parameter $P'_{u,v} = 0$. In an undirected network, the influence can be transmitted in both directions. The formula $P'_{u,v} > P'_{v,u} \neq 0$ when $C_u > C_v \neq 0$ is established in the adjusted model, too.

### B. CENTRALITY MEASUREMENT METHODS

Many centrality measurement methods have been proposed. Their categories and features are listed in Table 1 [22].




(a)




(b)

**FIGURE 7. The Influence Spreads of Seed Nodes Selected by Different Centrality Measurement Methods in GreedyCIC with Different λ Values under Different Datasets. For the NewGreedyIC algorithm, all the edge parameters are equal to λ. (a) Seedset size = 10, with GreedyCIC, under FBMIT. (b) Seedset size = 10, with GreedyCIC, under Epinions.**

The local measurement method only focuses on the features of the node itself. Taking degree centrality as an example, it takes the degree of each node as the centrality but does not consider the relationships between the node and other nodes in the same network. Global measurement evaluates the centrality of each node with respect to its relationships to the other nodes in the same network.

In this paper, the different centrality measurement methods can simulate the noise in influence diffusion models. We evaluate the performances of different types of centrality measurement methods in new algorithms.

The first centrality measurement method is degree centrality. It is defined as:

$$C_d(v_i) = \frac{d_{v_i}}{n-1}, \tag{8}$$

where $d_{v_i}$ is the degree of node $v_i$ and $n$ is the number of nodes in the network. For a directed network, $d_{v_i} = d_{v_i}^{in} + d_{v_i}^{out}$. Formula 8 is the normalized degree centrality with a normalization factor that is designed to avoid the influence of the volume of the network. Degree centrality is the most basic and simplest centrality measurement method for evaluating the centrality of some complex networks. We select it as one of the testing methods.

Another method is closeness centrality. It is defined as:

$$C_c(v_i) = \frac{1}{\overrightarrow{l_{v_i}}}, \quad \overrightarrow{l_{v_i}} = \frac{1}{n-1}\sum_{i\neq j}\overrightarrow{l_{v_i v_j}}, \tag{9}$$

where $\overrightarrow{l_{v_i}}$ is the average shortest distance between node $v_i$ and all the other nodes in the same network, and $\overrightarrow{l_{v_i v_j}}$ is the shortest distance between node $v_i$ and node $v_j$. Closeness centrality was first defined by Bavelas in 1950 as the reciprocal of farness, i.e., $\overrightarrow{l_{v_i}}$. In this method, a node has higher closeness centrality when it becomes closer to the other nodes.

The third method is eccentric centrality. It is defined as:

$$C_e(v_i) = \frac{1}{dist_{max}(v_i)}. \tag{10}$$

Eccentricity is a term in graph theory. It describes the degree of how far the node is from the centre of the graph. In centrality measurement, it is the largest shortest path among all the shortest paths from the node to the other nodes in the same network.

The last method is PageRank centrality. It is defined as:

$$C_p(v_i) = \alpha \sum_{j=1}^{n} A_{i,j} \frac{C(v_i)}{d_j^{out}} + \beta. \quad (11)$$

PageRank is a classic algorithm that was first proposed by Google to address the webpage searching problem. It defines a PageRank value such that a node with high PageRank value has high centrality value and it evaluates the outdegree of the node in a directed network.

These four centrality measurement methods evaluate the centrality from various aspects. Other methods, such as Katz centrality and Eigenvector centrality, are similar to degree centrality and PageRank centrality. Therefore, the methods we selected above could represent the general centrality measurement methods, which could be an important noise factor in robust influence maximization.

### C. ITERATION PROCESS IMPROVEMENT AND ITERATION STOPPING CONDITION

Kempe *et al.* have already proved that the *influence maximization problem* is an NP-hard problem. It is NP-hard to find the seed set with $k$ nodes. Since the influence function $\sigma_\theta(S)$ in Formula 1 has been proved submodular, Kempe gives a $(1-\frac{1}{e})$ approximation using a general greedy method in the *Independent Cascade Model* [1].

In this paper, we use the basic principle of the robust influence maximization objectives proposed by He that $\rho(S) = \min_{\sigma \in S} \frac{\sigma(S)}{\sigma(S^*_\sigma)}$. Then, in the CIC model, we add the node centrality and edge parameter as the factors during influence propagation, so the **new Robust Influence Maximization Problem** is defined as in Formula (12).

For a given graph $G = (V, E)$, the centrality space $C$ of all the vertices and the edge parameter space $P$ of all the edges are defined. With the size of the seed set $k$, we are required to find a seed set $S \subseteq V$ of $k$ vertices that maximizes the **robust influence maximization objective**:

$$S^*_{C,P} := \frac{\arg\max}{S \subseteq V, |S| = k} \left( \frac{\min}{c \in C, p \in P} \frac{\sigma_{c,p}(S)}{\sigma_{c,p}\left(S^*_{c,p}\right)} \right). \quad (12)$$

This problem aims to find the seed set $S^*_{C,P}$ that has the largest robust objective. The two spaces of centrality and edge parameter are the noise in the influence model. Different results of centrality measurement methods compose the space $C$. The space $P$ contains edge parameters calculated by different centrality values [1].

The specific problem is to find appropriate centrality measurement algorithms to improve the accuracy and efficiency of influence maximization algorithms. According to the different results of the algorithms, we will find the best parameters and algorithms that maximize the robust influence maximization objectives.

## IV. ALGORITHM FRAMEWORK

There are two basic algorithms for *influence maximization problems*: the greedy method and heuristic method. Kempe *et al.* propose the Greedy Hill Climbing algorithm and heuristic algorithms based on degree, centrality and a random method [1].

The greedy algorithm has high accuracy but low efficiency. It uses the greedy method to traverse all the nodes in the graph and calculate the marginal utility. For large-scale social graphs, the time cost of traversal is very high.

For heuristic algorithms, Kempe *et al.* choose the initial nodes according to node centrality or node degree, or just randomly select $k$ nodes. This reduces the running time but has low accuracy compared with greedy methods. Some papers improve the original greedy algorithms and heuristic algorithms greatly and take the research of influence maximization to a new level [2], [3], [6]. It is viable to solve *influence maximization problems* on large-scale networks.

In this paper, we study two classic algorithms: the NewGreedyIC and DegreeDiscount proposed by Chen in 2009 [2]. These two algorithms outperform the previous algorithms greatly in running efficiency, especially the heuristic algorithm.

The NewGreedyIC improved upon the original greedy algorithm proposed by Kempe. It added the edge probability as one of the inputs of the new algorithm. The default parameter value is $p = 0.01$. For calculating the marginal utility, it uses the Breadth First Search (BFS) method. Comparing with the CELF optimization proposed by Leskovec *et al.* [6], it has 15% to 34% lower running time than the CELF method. The NewGreedyIC algorithm has the same influence spread as the original greedy algorithm.

In this paper, we add the edge parameter space $P$ calculated by various centrality measurement methods as the input of the NewGreedyIC algorithm and our new algorithm is called GreedyCIC.

The DegreeDiscount algorithm improves upon the former DegreeHeuristic algorithm with a degree-discount method. For each node $v$ in a network, the algorithm calculates the degree discount of each node and finally chooses the $k$ nodes with the largest degree discounts as the seed set.

$$dd_v = d_v - 2t_v - (d_v - t_v)\, t_v p \quad (13)$$

Formula 13 defines the degree discount. $t_v$ is the number of neighbours of vertex $v$ that have already been selected as seeds. The basic idea is that when we select node $u$ as the seed, if node $v$, which is the neighbour of $u$, is already in the seed set, we should not consider the edge between $u$ and $v$ in the degree calculation of node $u$. Therefore, Formula 13 is proposed to eliminate the influence. The new heuristic algorithm has been proved to have better influence spread than the previous degree heuristic and centrality heuristic methods.

In this paper, we also change the input $P$ of the DegreeDiscount algorithm to the edge parameter spaces calculated by the different centrality measurement methods. Our new algorithm for that is called NewDiscount. We change the degree-discount definition to Formula 14:

$$dd'_v = d_v - 2t_v - (d_v - t_v)\, t_v p\, (u, v)\,. \tag{14}$$

The former $p$ is set to the default value of 0.01. From the edge parameter space, we can apply the exact parameter value to each edge from node $u$ to node $v$. To show the procedure of the algorithms clearly, Figure 3 is designed to illustrate the logical process of the algorithms proposed by us.

## V. EXPERIMENT

### A. EXPERIMENTAL SETUP

For computing the node centrality, we use the Python package called Snap.py from Stanford [24]. All the codes are run under the Ubuntu 16.04 LTS with Quad-Core Intel Xeon E5-2407 v2 2.4 GHz and 40 GB memory. To investigate the robustness of the influence propagation models and algorithms, we perform different types of experiments with different algorithms, models and datasets. For the same algorithm, different sets of coefficient values also represent the noise in the model. Then, we compare the results of experiments vertically and horizontally.

In our experiment, we mainly run the following algorithms in various experimental setups.

**Algorithm 1** GreedyCIC ($G$, $k$, $P$)

set $S = \emptyset$ and $R = 200$
for $i = 1$ to $k$ do
  set $s_v = 0$ for all $v \in V \backslash S$
  for $i = 1$ to $R$ do
    compute $G'$ by removing each edge from $G$ with probability $1 - p$ where $p \in$ P
    compute $R_{G'}(S)$
    compute $|R_{G'}(\{v\})|$ for all $v \in V$
    for each vertex $v \in V \backslash S$ do
      if $vR_{G'}(S)$ then
        $s_v += |R_{G'}(\{v\})|$
      end if
    end for
  end for
  set $s_v = s_v/R$ for all $v \in V \backslash S$
  $S = S \cup \{\text{argmax}_{v \in V \backslash S}\{s_v\}\}$
end for
output $S$

1) **GreedyCIC**: This greedy algorithm (Algorithm 1) is modified from the NewGeedyIC algorithms proposed by Chen et al. We add edge parameter space calculated by centrality to the model. We set the R=200 as the default value.

2) **NewDiscount**: This is the heuristic algorithm (Algorithm 2) improved from DegreeDiscount proposed by Chen et al. We add the parameter space and improve the method of calculating the degree discount.

**Algorithm 2** NewDiscount ($G$, $k$, $P$)

set $S = \emptyset$
for node $v \in$ G do
  compute degree $d_v$
  $dd_v = d_v$
  set $t_v = 0$
end for
for $i = 1$ to $k$ do
  select $u = \text{argmax}_v\{dd_v | v \in V \backslash S\}$
  $S = S\ \{u\}$
  for each neighbour $v$ of $u$ and $v \in V \backslash S$, $p \in P$ do
    $t_v = t_v + 1$
    $dd_v = d_v - 2t_v - (d_v - t_v)t_v p\ (u, v)$
  end for
end for
output $S$

3) **NewGreedyIC**: This algorithm is proposed by Chen et al. [2]. We set R=200. The algorithm is set for comparison.

4) **DegreeDiscount**: This algorithm is proposed by Chen et al. [2]. The algorithm is set for comparison.

To produce the robustness results, we use the robust influence maximization (RIM) optimization objective in Formula 2 proposed by He and Kempe [15]. Based on the experimental results, we will evaluate the RIM optimization objective and choose the algorithm that has the most robust performance.

### B. DATASETS

Since our paper focuses on online social networks, all the datasets we choose are collected from online social networks. Furthermore, to address the noise in the influence propagation model, we choose various datasets from different online social networks with different sizes. Some of the networks are directed networks. Some of the networks have the feature that nodes are more likely to be influenced by other nodes. For example, the Twitter retweeting network has this feature, but the scientist collaboration network does not have this feature since the selection of co-authors is not related to the influence of the authors themselves. The following are the datasets we utilize in our experiment. Details of these datasets can be found in Table 2.

1) **Retweet**: It is a dataset collected from the online microblog website Twitter. It contains the nodes that represent all the users in the retweeting network; edges represent all the retweets among the users. The retweeting information is collected from various hashtags, including #political and #copen, which represent The United Nations Climate Change Conference hosted in Copenhagen in 2012. The retweet dataset has the feature that users are more likely to retweet the tweets posted by more influential users and the influences of users are related to their centralities [25].

2) **FBMIT**: It is a dataset collected from the online social website Facebook. The dataset contains users from one hundred universities and colleges in America. We choose the subset in MIT, which contains Facebook users studying at the Massachusetts Institute of Technology. This dataset does not have the feature described above because the connections between users are closer and one user cannot be easily influenced by others [26].

3) **Epinions**: This is crawled from a user-oriented online product reviewing website, www.epinions.com. On this website, users are encouraged to select the users they trust. Therefore, this dataset is from a web of trust, and the edges in the network represent one user's trust in another user. This dataset also has the same feature as the Retweet dataset, that users with higher centrality are more influential [27].

4) **Douban**: Douban.com, which was launched on March 6, 2005, is a Chinese Web 2.0 website that provides user review and recommendation services for movies, books, and music. It is also the largest online Chinese language book, movie and music database and one of the largest online communities in China. This dataset contains the friendship network that was crawled in December 2010 by Long Qiu (lqiu4@asu.edu). This dataset does not have the same feature as the Retweet dataset [28].

### C. EXPERIMENTAL RESULTS AND DISCUSSION

The edge parameter is calculated from the centralities of nodes. There are four different centrality measurement methods in the CIC model. The first experiment is to compare the performances of different centrality measurement methods in estimating the edge parameters.

First, we investigate the performances of different centrality measurement methods in setting the edge parameter space under the NewDiscount algorithm with different values of the seed set size $k$ and different datasets. We set the seed set size from 1 to 46 and use the influence propagation algorithm in the *Independent Cascade Model* to simulate the influence diffusion from seed nodes that we selected with the NewDiscount algorithm. For the influence propagation algorithm, we simulate the influence propagation process 1000 times and take the average influenced node set size as the final number of influenced nodes (i.e., the y-axis).

From Figure 4, we find that Degree Centrality and PageRank Centrality perform better than the other two methods on all the datasets for different seed set sizes. When the size of dataset becomes larger, the differences in performance between the first two methods and last two methods become larger. The first two centrality measurement methods have almost have the same performance when the seed set size is 1. When the seed set size is small, Degree Centrality performs better than Page Centrality; when the seed set size increases, PageRank Centrality outperforms the Degree Centrality.

The Eccentric Centrality measurement methods have the worst performance under this experiment set. The Eccentric Centrality depends only on the distance from the centre, and the relationships between the nodes and network structure are not considered.

This may be the reason for its bad performance. PageRank Centrality has the best performance. This shows that PageRank Centrality can simulate the real influence better than other methods in the online social network.

Another interesting phenomenon is that the gradient of Influence Spread in (b) approaches zero faster than in other charts. From Table 2, we find that the FBMIT dataset has a much larger clustering coefficient than the other three datasets, which indicates that the nodes in FBMIT are closer than the nodes in other datasets. Thus, influence propagates in FBMIT faster than in other datasets.

To investigate the differences among various centrality measurement methods in selecting seed sets, we take the Retweet dataset as an example and use the visualization tool *Gephi* to show the seed nodes selected with different methods. Because Retweet has small numbers of nodes and edges, we can see the relationship among the nodes clearly. We set the size of the seed set to 15.

In Figure 5, we see that almost 80% of the nodes selected by the four methods are the same nodes. This shows that all four methods have good performances on the small datasets. Comparing Figure 5 (a) with (b), (c) and (d), (c) and (d) select the nodes that influence more nodes and do not select the nodes that influence fewer nodes. This indicates that the results of the algorithms are not robust. We cannot rely on the results calculated by only one method.

**TABLE 3. Average running times (ARTs) of different algorithms in different experimental datasets.**

| Algorithm | Method | Dataset | ART (seconds) |
|---|---|---|---|
| Degree | Heuristic | FBMIT | 22.2 |
| | | Epinions | 30.1 |
| PageRank | Heuristic | FBMIT | 24.1 |
| | | Epinions | 29.9 |
| Eccentric | Heuristic | FBMIT | 22.5 |
| | | Epinions | 36.9 |
| Closeness | Heuristic | FBMIT | 24.6 |
| | | Epinions | 42.4 |
| DegreeDiscount | Heuristic | FBMIT | 22.9 |
| | | Epinions | 27.0 |
| Degree | Greedy | FBMIT | 268.7 |
| | | Epinions | 839.7 |
| PageRank | Greedy | FBMIT | 203.0 |
| | | Epinions | 842.0 |
| Eccentric | Greedy | FBMIT | 202.6 |
| | | Epinions | 880.2 |
| Closeness | Greedy | FBMIT | 198.4 |
| | | Epinions | 881.2 |
| NewGreedyIC | Greedy | FBMIT | 194.7 |
| | | Epinions | 905.1 |

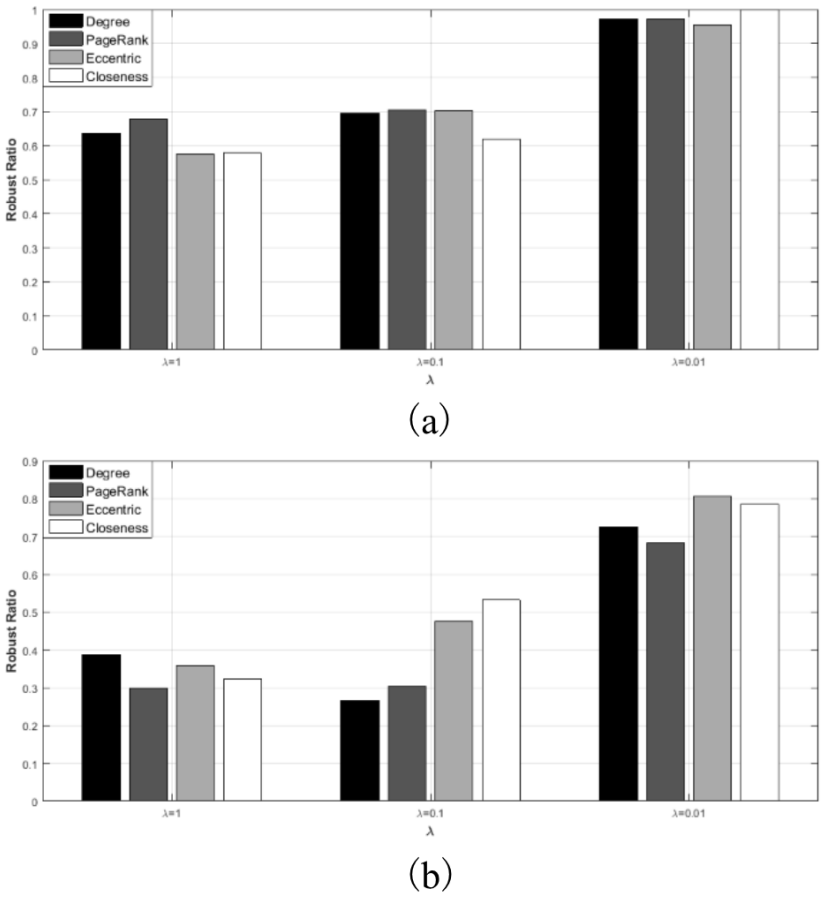



**FIGURE 8.** Robustness Performances of Different Centrality Measurement Methods under Different Modification Coefficients in Datasets (a) FBMIT and (b) Epinions. All the edge parameters are equal to λ.

According to the influence spread results shown in Figure 4, PageRank Centrality has the best performance. It is also observed in Figure 5 (b) that the seed nodes connect more nodes. Then, we investigate the influence of the modification coefficient in Formula 7. We run the NewDiscount algorithm for a seed set size of 10 on the FBMIT dataset.

In Figure 6, we see that Degree Centrality and PageRank Centrality perform similarly to the original algorithm. DegreeDiscount slightly outperforms the other two algorithms in magnitude. Since the edge parameter is set to a default value of 0.01 in the original DegreeDiscount algorithm, we find that by setting $\lambda = 0.01$, we can simulate the original influence spread well.

For the same experiment with the GreedyCIC and NewGreedyIC algorithms, the results are shown in Figure 7. Degree Centrality and PageRank Centrality have almost the same performance as NewGreedyIC in FBMIT, but PageRank has better performance than NewGreedyIC on all sets of seed nodes. When dataset becomes larger, the new algorithms based on the CIC model outperform the original algorithm for the greedy method.

Comparing the results of the heuristic and the greedy methods, when dataset size is larger, the performances of the heuristic algorithms are better than those of the greedy algorithms. However, the greedy algorithms should have greater influence spreads than the heuristic algorithms. The reason is that we set the number of rounds of the greedy algorithms to 200, which is much less than the default value of 20000. According to the conclusion of [2], increasing the number of rounds R will not improve the performance of the greedy algorithm. The experimental result shows that this conclusion has the indispensable precondition that R is large enough.

Incorporating centrality measurement methods into the traditional *influence maximization problem* will increase the running times of the algorithms. Table 3 shows the running times of different algorithms. For greedy algorithms, the new algorithms have lower run times than the original algorithm. However, for the heuristic algorithms, the original algorithm has a lower run time.

To investigate the robustness of the algorithms, we take the robust optimization objective of [14] and investigate the robustness performances of different centrality measurement methods under different modification coefficient values. The robustness performance is defined by Formula 2 as the minimum ratio of the influence spreads of all seed sets to the best influence spread among all the influence propagation

functions. From Figure 8, all the methods have the best performance when $\lambda = 0.01$. For different centrality measurement methods, the Eccentric Centrality has the most robust performance on all the experiment sets.

## VI. CONCLUSION & FUTURE WORK

In this paper, we incorporate the centrality measurement methods into the process of determining the edge activation probability and propose a Centrality-based Independent Cascade (CIC) model. Under this model, we improve the DegreeDiscount and NewGreedyIC algorithms by adding a centrality-based edge parameter space to them and propose two algorithms: the NewDiscount and GreedyCIC algorithms.

According to the experimental results on four datasets, parameter spaces generated by Degree Centrality and PageRank Centrality have wider influence spreads than those generated by the other two methods. As the dataset size grows, the performance of PageRank Centrality gradually exceeds the performance of Degree Centrality. For selecting the seed set, all the algorithms could select most of the influential nodes and the differences among different centrality measurement methods are not obvious. This indicates that noise exists in the algorithms and could influence the performances of the algorithms.

The modification coefficient for measuring the edge parameter is set as the noise in the algorithms. According to the experimental results under FBMIT and Epinions, Degree Centrality and PageRank Centrality could simulate the original algorithms better for all sets of modification coefficients. PageRank centrality in GreedyCIC performs better than NewGreedyIC on Epinions for all coefficient value sets. We predict that the PageRank-based GreedyCIC algorithm will outperform the original NewGreedyIC when the size of the dataset grows.

We define the robustness performance according to the former optimization objective and evaluate it under different modification coefficient values. The results show that the algorithms under the CIC model have more robust performances when the modification coefficient set to 0.01 and the Eccentric-Centrality-based algorithms have more robust performances among the four algorithms based on different centrality measurement methods. The Eccentric-Centrality-based algorithms have the worst influence spreads among all the algorithms, but the best robustness performances. This indicates that the robustness and performance are two different aspects of influence maximization algorithms; it seems hard to optimize both aspects at the same time.

To summarize, incorporating centrality measurement methods into the process of measuring the edge parameter could improve the influence spread in some situations. The PageRank-Centrality-based algorithms have the best performances in terms of influence spread and running time among all four centrality measurement methods. In the CIC model, algorithms have different robustness performances under different values of the modification coefficient, and they attain the best robustness performances when the modification coefficient is set to 0.01.

For noise in the algorithms, more noise should be considered in further research. Different models such as the *Linear Threshold Model*, different algorithms such as CELF optimization and topic-aware algorithms, and different datasets such as the Weibo dataset with millions of nodes should be added to the experiments to evaluate the robustness of the algorithms, and different sizes of experimental datasets should be investigated.

For the Centrality *Independent Cascade Model*, additional centrality measurements should be added to find methods that perform better than PageRank. As we supposed above, an online social network with the feature that users are more likely to be influenced by others may have better performance when a centrality measurement method is added. To prove this hypothesis, we need to perform controlled experiments in further research.

## REFERENCES

[1] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. KDD*, 2003, pp. 137–146, doi: 10.1145/956750.956769.

[2] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. KDD*, 2009, pp. 199–208, doi: 10.1145/1557019.1557047.

[3] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. ICDM*, Dec. 2010. pp. 88–97, doi: 10.1109/ICDM.2010.118.

[4] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. WWW*, 2011, pp. 47–48, doi: 10.1145/1963192.1963217.

[5] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. SODA*, 2014, pp. 946–957, doi: 10.1137/1.9781611973402.70.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. KDD*, 2007, pp. 420–429, doi: 10.1145/1281192.1281239.

[7] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," *Knowl. Inf. Syst.*, vol. 37, no. 3, pp. 555–584, 2013, doi: 10.1007/s10115-013-0646-6.

[8] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-K influential nodes in mobile social networks," in *Proc. KDD*, 2010, pp. 1039–1048, doi: 10.1145/1835804.1835935.

[9] N. Du *et al.*, "Scalable influence estimation in continuous-time diffusion networks," *Adv. Neural Inf. Process. Syst.*, vol. 26, no. 2, pp. 3147–3155, 2013.

[10] P. Netrapalli and S. Sanghavi, "Learning the graph of epidemic cascades," in *Proc. SIGMETRICS*, 2012, pp. 211–222, doi: 10.1145/2318857.2254783.

[11] M. Gomez-Rodriguez *et al.*, "Uncovering the temporal dynamics of diffusion networks," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 561–568.

[12] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Knowledge-Based Intelligent Information and Engineering Systems—KES*, 2008, doi: 10.1007/978-3-540-85567-5_9.

[13] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. WSDM*, 2010, pp. 241–250, doi: 10.1145/1718487.1718518.

[14] X. He and D. Kempe, "Robust influence maximization," in *Proc. KDD*, 2016, pp. 885–894, doi: 10.1145/2939672.2939760.

[15] X. He and D. Kempe. (Jan. 2015). "Stability of influence maximization." [Online]. Available: https://arxiv.org/abs/1501.04579

[16] W. Chen, T. Lin, Z. Tan, M. Zhao, and X. Zhou, "Robust influence maximization," in *Proc. KDD*, 2016, pp. 795–804, doi: 10.1145/2939672.2939745.

[17] M. Lowalekar, P. Varakantham, and A. Kumar, "Robust influence maximization," in *Proc. Int. Conf. Auto. Agents Multiagent Syst., Int. Found. Auto. Agents Multiagent Syst.*, 2016, pp. 1395–1396.
[18] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "SIMPATH: An efficient algorithm for influence maximization under the linear threshold model," in *Proc. ICDM*, Dec. 2011, pp. 211–220, doi: 10.1109/ICDM.2011.132.
[19] X. He, G. Song, W. Chen, and Q. Jiang, "Influence blocking maximization in social networks under the competitive linear threshold model," in *Proc. SDM*, 2012, pp. 463–474, doi: doic 10.1137/1.9781611972825.40.
[20] Y.-C. Chen, W.-Y. Zhu, W.-C. Peng, W.-C. Lee, and S.-Y. Lee, "CIM: Community-based influence maximization in social networks," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 2, 2014, Art. no. 25, doi: 10.1145/2532549.
[21] Alongyin. *The Research of IC Model and it Model*. [Online]. Available: http://blog.csdn.net/hit090420216/article/details/44755335
[22] M. E. J. Newman, *Networks: An Introduction*, vol. 327. Oxford, U.K.: Oxford Univ. Press, 2010.
[23] A. Bavelas, "Communication patterns in task-oriented groups," *J. Acoust. Soc. Amer.*, vol. 22, no. 6, pp. 725–730, 1950, doi: 10.1121/1.1906679.
[24] J. Leskovec and R. Sosic. (Jun. 2016). "SNAP: A general purpose network analysis and graph mining library." [Online]. Available: https://arxiv.org/abs/1606.07550
[25] R. A. Rossi, D. F. Gleich, A. H. Gebremedhin, and M. M. A. Patwary, "What if CLIQUE were fast? Maximum cliques in information networks and strong components in temporal networks," pp. 1–11, 2012. [Online]. Available: http//arXiv:1210.5802
[26] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *Phys. A, Statist. Mech. Appl.*, vol. 391, no. 16, pp. 4165–4180, Feb. 2011. [Online]. Available: https://arxiv.org/abs/1102.2166
[27] M. Richardson, R. Agrawal, and P. Domingos, *Trust Management for the Semantic Web*. Berlin, Germany: Springer, 2003, doi: 10.1007/b14287.
[28] H. L. R. Zafarani. *Social Computing Data Repository at ASU*. [Online]. Available: http://socialcomputing.asu.edu
[29] P. S. Aric and S. H. Dan. *Networkx: High-Productivity Software for Complex Networks*. Accessed: Sep. 2017. [Online]. Available: http://www3.cs.stonybrook.edu/~algorith/implement/networkx/implement.shtml
[30] S. Yu, M. Liu, W. Dou, X. Liu, and S. Zhou, "Networking for big data: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 531–549, 1st Quart. 2017.
[31] L. Gao, T. H. Luan, S. Yu, W. Zhou, and B. Liu, "FogRoute: DTN-based data dissemination model in fog computing," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 225–235, Feb. 2017, doi: 10.1109/JIOT.2016.2645559.
[32] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016, doi: 10.1109/ACCESS.2016.2577036.
[33] S. Yu, G. Wang, and W. Zhou, "Modeling malicious activities in cyber space," *IEEE Netw.*, vol. 29, no. 6, pp. 83–87, Nov./Dec. 2015, doi: 10.1109/MNET.2015.7340429.

**XIAOLONG DENG** was born in 1977. He received the Ph.D. degree. He is currently an Assistant Professor and a Master Supervisor in data mining with the Beijing University of Posts and Telecommunications. His research interests include data mining and complex networks.

**YINGTONG DOU** was born in 1995. He received the bachelor's degree. He is currently a Senior Student with the Beijing University of Posts and Telecommunications and the Queen Mary University of London. His research interests include social network analysis and data mining.

**TIEJUN LV** (SM'12) was born in 1969. He received the Ph.D. degree. He is currently a Professor with the Beijing University of Posts and Telecommunications. His research interests include network communication and signal processing.

**QUOC VIET HUNG NGUYEN** received the Ph.D. degree. He is currently a Lecturer with Griffith University. He has published several papers in conferences and journals, such as SIGMOD, ICDE, IJCAI, and TKDE. His research interests include data exploration.