DIN : An Efficient Algorithm for Detecting Influential Nodes in Social Graphs using Network Structure and Attributes

Myriam Jaouadi
MARS Research Unit, SDM Research Group
ISITCom, University of Sousse, Tunisia
Email: jaouadimaryem@gmail.com
Telephone: +216 53073323

Lotfi Ben Romdhane
MARS Research Unit, SDM Research Group
ISITCom, University of Sousse, Tunisia
Email: lotfi.ben.romdhane@gmail.com

***Abstract*—Detecting influential nodes in social networks represents an essential issue for various applications to identify users that may maximize the influence of information in such networks. Several methods have been proposed to solve this problem often khown as influence maximization problem. However, most of them focused on the structure of the network and ignored the semantic aspect. Besides, these methods are parametric, they require the number k of influential elements in a deterministic manner. In this paper, we propose a parameterless algorithm called DIN (Detecting Influential Nodes in social networks) that combines the structure and the semantic aspect. The main idea of our proposal is to detect communities with overlap, modelize the semantic of each community then select influential elements. Experimental results on computer-generated artificial graphs demonstrate that DIN is efficient for identifying influential nodes , compared with two newly known proposals.**

1. Introduction

Social networks have largely changed the way we produce, distribute and consume information and are therefore become carriers of important informations. Social network analysis has received increasing interest in many different areas in recent years, including community detection, role detection, etc [4]. In this paper, we adress one of the most important fields in social network analysis, namely detection of influential nodes in a social network.

The detection of most influential people has attracted many researchers in the field of social networks. This problem is also known as influence maximization (IM) and has many applications such as: opinion propagation, studying acceptance of political movements or acceptance of technology in economics [1]. For example a customer that would like to advertise a novel product can present free samples to an initial number of individuals hoping that they try to convince their friends to buy it and the recommendation circulates among friends and friends of friends.

Influence maximization is an NP-hard (non-deterministic polynomial-time) problem depending on the information diffusion model in social networks [2]. Two well-known diffusion models were presented for the first time in [3]: Linear Threshold (LT) and Independent Cascade (IC). In this paper, inspired by LTM, we define a novel information diffusion model for controlling the propagation of information through a set of candidate nodes. We start by the detection of overlapping communities, then we model the semantics of each one based on interests of its members and the candidate nodes are generated based on the PageRank [17] algorithm . Finally, we apply our diffusion model for the selection of influential nodes such that they can spread information to the largest number of nodes in the network.

The remainder of this paper is organized as follows: Section 2 surveys related work. Section 3 states the problem we address. Section 4 describes the fundamentals of our model. Section 5 reports experimental results on artificial networks and the final section concludes this paper and sheds light on future research directions.

2. Related work

Influence maximization problem is an issue of a great importance in social networks analysis. The works of [3], [5] were the first that provided a formulation for influence maximization as an optimization problem. Many other works have been done, we propose to categorize them into six classes that study, on the one hand, the structure of the network and on the other hand, the activity of the network members.

A. Communities detection-based approaches

This family of approaches is based on the structure of the network. The goal of communities detection is to search strongly linked groups of vertices. It can provide a simplified representation of the structure of large scale networks. In fact, it facilitates the identification of influential nodes, thus while searching in the whole network, we can search in each community independely and reduce the research space. CIM (Community-Based Influence Maximization) [6] is a community based algorithm for influence maximization problem that starts by the detection of overlapping communities. Then, transforms the unweighted graph to weighted one by the measure of similarity among nodes and identifies the centroid

978-1-5090-4320-0/16/$31.00 ©2016 IEEE

nodes of each community as candidates. The last phase of this framework is to finalize the seed nodes from the candidate set with consideration that the size of community is still a factor for seed placement. In fact, it relies on the largest communities for seed selection. ComPath [1] is an approach that identifies community structures of the input graph to form a new network. Then, the most influential communities are selected among them based on the betweeness centrality. Next, a number of nodes is chosen from each community based on the degree centrality to form the set of candidate nodes. Finally, ComPath selects top-k most influential nodes from the candidates based on the paths that exist in the network. Bae et al. [18] presented a centrality measurement called coreness centrality based on the $k-shell$ decomposition to quantify the ability of propagation of a node using centrality information of its neighbors. It should be noted that high degree nodes are not always most influential in all contexts. Walia et al. [9] have proposed an algorithm that, first, detects communities with overlap based on LFM (local maxima of a fitness function) method [10]. Then, finds the importance of overlapping nodes at each community they belong to and finalizes by the choice of the most important overlapping nodes that maximize an objective function as seeds. This method depends on a parameter that controls the size of formed groups [11] and the complexity in the worst case is $O(n^2)$ where n is the number of nodes. This family of approaches is based on the structure information with ignorance of the semantic aspect of the network.

B. PageRank-based approaches

PageRank algorithm [17] is one of the most popular web algorithms. It was used by Google search engine for links analysis. The main idea of PageRank is to use link information to measure sites importance and thus be able to classify the result of information retrieval. FBI (Fine-Grained Feature-Based Social Influence Evaluation) [13] is an algorithm that evaluates a user's social influence on the basis of its features including related topics or user profiles. FBI starts by the construction of initial influence that combines the importance of the user its self and its effect on his neighbors. Then, the initial influence is adjusted based on the PageRank algorithm to construct the social influence. The time complexity of FBI is $O(a^2m)$, where a is the average degree of nodes, and m is the number of nodes. Therefore, it can expect high complexity with large-scale networks.

C. Greedy algorithm-based approaches

The main idea of this family of approaches is to select the best individual that maximizes the spread of information. Then, select the second one that, by adding it to the chosen node, it can increase the marginal gain of a certain objective function. A greedy algorithm iterates k steps to find top-k influential nodes and in each iteration the most influential node is found and added to the seed set [7]. To reduce computations, CELF++ (Cost Effective Lazy Forward) [8] avoids unnecessary recalculations of marginal gains incurred by a naive greedy algorithm. The idea of CELF++ is that the marginal gain of a new node is reduced when the number of seed nodes increases. It uses a data structure that stores next step marginal gain for a particular node. Although this approach optimizes the greedy algorithm, it increases memory consumption due to Monte-Carlo simulation and could not considerably reduce recalculations of propagation. SMG (State Machine Greedy) [7] is another approach that changes the way Monte-Carlo simulation is calculed by preventing recreation of Monte-Carlo graph instances. In fact, it saves calculated propagation of seed nodes in such instances as last state and uses it to prevent recalculation in the next step. Then, to simulate the diffusion process and estimate the marginal gain of each candidate, SMG traverses only nodes of the last state contrary to the greedy algorithm that traverses the whole graph. This family, even it reduces calculations, but it is still time consuming.

D. EM algorithm-based approaches

The standard procedure for maximum likelihood estimation of latent variable model is Expectation Maximization (EM) algorithm [14]. Several works have focused on such algorithm to estimate parameters with maximum likelihood. TSIM (Topic-sensitive Influencer Mining) [14] is a model that aims to extract the influential users from social networks images based on topics of shared images, specifically, it takes Flickr as the study platform. TSIM Constructs an hypergraph in which we find two types of nodes (users and images), extracts informative images and studies topics from the extracted ones. Then, evaluates a score for users and images based on EM algorithm. Finally, it classifies sensitive influencers for a specific topic k. Wang et al. [15] proposed a model to verify the existence of the influence of emotion in image networks. The influence of emotion means that the emotional state of a user is influenced by his friends through their published images [15]. It includes contextuel and visual aspects of images. However, this family neglects information diffusion models.

E. Linear threshold and Indepedant Cascade-based influence maximization algorithms

Linear Threshold Model (LTM) and Independant Cascade Model (ICM) are the most popular information diffusion models. Both LTM and ICM are stochastic models in which information flows from a node to its neighboring nodes at each time-step according to some probabilistic rule [4]. Several approaches tried to optimize those models. CTMC-ICM (Continuous-Time Markov Chain-Independant Cascade Model) [4] is an approach derived from the ICM, which effectively calculates a good estimation of influential nodes. This method improved the ICM by incorporation of the theory of Continuous-Time Markov Chain (CTMC). PSI (Probabilistic Social Influence model) [16] is a model that combines the best aspects of ICM and LTM with elimination of their limitations to introduce a probabilistic distribution of social influence. The basic idea of such model is, instead of using a uniform diffusion probability, PSI defines an activity based probability for each node. Although, it is simple to capture the basic

principle of both models, they are not useful for predicting the behavior of the network.

F. Agent-based approaches

Finally, the agent-based approach TWC-TSS (Time Window Constraint-Target Set Selection) [12] is a mathematical framework that uses agents behaviours for the selection of influential nodes in a limited time interval whereas agents have an unbounded memory. In fact, it studies the information propagation in networks where agents change their opinions/behaviours on the basis of the behaviours of their neighbors.

Inspired by the above mentioned approaches, we define a parameterless approach that combines the structure and semantic aspect of the network to identify most influential nodes. Before detailing our proposal, we need to introduce some basic concepts.

3. Preliminaries

A. Problem formulation

We consider an undirected graph $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges. The purpose of the detection of influential nodes in G is to determine users that result in a maximum diffusion of information. We define a new method for detecting influential elements and we propose a new diffusion model. Our method covers a significant number of nodes in the graph and does not require a priori knowledge of the number of influential elements to detect. According to the proposed diffusion model for which we seek to maximize the influence, we will need additional parameters such as users' interests on which we will concentrate for modeling the semantic aspect of the network.

B. Basic defenitions

Definition 1: (Community)
We adopt the popular definition of a community that is a set of dense nodes having more internal links than external [11].

Definition 2: (Overlapping nodes)
overlapping nodes designate the nodes that belong to more than one community at a time.

Definition 3: (Network semantics)
The semantic of a social network is any information that characterizes network users such as profile information, related topics, their interests, communication, actions (like, share, comment ...).
For our proposal, we will concentrate on users' interests for modeling the social network semantics.

Definition 4: (User interests)
Interests designate the passions (such as reading, music, sport) of a user.
More formally, we note by $U = (U_1, U_2, ..., U_N)$ the set of N users of G. For each user, we will associate a set of interests that characterize it. These interests will be presented as a characteristic vector.

Definition 5: (A user's characteristic vector)
Each user U is represented by a characteristic D-dimensional vector $X_i = (x_{i1}, ..., x_{iD})$ where each element corresponds to the interest x_d of the user U_i. The size of this vector varies from one user to another.

Definition 6: (Active node)
A node is said to be active if it adopts an idea or innovation and diffuses it, in other terms is influenced by a content diffused in the network.

Definition 7: (Area of influence)
The area of influence of a node u that we note $\sigma(u)$ can be determined by the set of nodes that u succeded to influence.

4. Our proposal: DIN (Detecting Influential Nodes in social networks

A. General introduction

Our proposal combines both structure and semantic aspect of the network to detect elements maximizing the influence (the influential nodes). For this reason, the main idea is to propose a two-phase approach:
1. Partition phase.
2. Selection phase.
The first phase of our model exploits the structural aspect of the network, while the second one exploits the semantic aspect.

Figure 1 describes the general principle of the proposed approach.

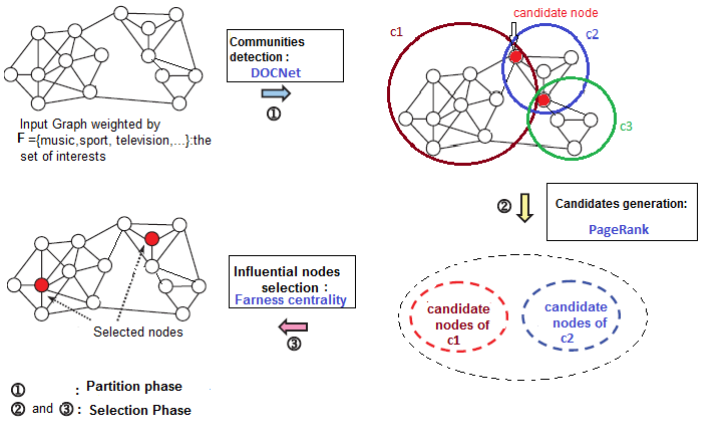

Fig. 1. General principle of the proposed approach

Given a graph G whose nodes are weighted by their interests, we start by partitioning the graph into communities which verifies our exploitation of the structural aspect of the network. This first phase applies a model of detection of overlapping communities. For this objectif, we propose to use the DOCNet model (Detection of Overlapping Communities in Networks) [11], which aims at detecting these communities, and provides the overlapping nodes that play an important role in information transfer. For more details please visit [11].

Then, we move to influential nodes selection phase. Based on users' interests, we model the semantic of our network. For this, we define new measures affecting a weight to each node and each edge of the graph. These measurements verify the

importance of a node's interests within its community and treat the similarity between the members of the same community.

1) Modeling semantics: Each community admits a defenite set f of interests characterizing its members; $C_k(x_{k1}, ..., x_{kf})$ where x_{ki} is the interest i $(i = 1, ..., f)$ in the community C_k. Therefore, we propose a frequency measurement called interest frequency of each interest as follows.

Definition 8: (interest frequency)
The frequency of an interest x_i in a community C_k is a measure to determine the number of users that share the same interest x_i in a specific community C_k. It can be defined by:

$$if(x_{ki}) = |U_j / x_{ki} \in X_j \text{ and } U_j \in C_k|. \quad (1)$$

where X_j is the characteristic vector of the user U_j.

Definition 9: (Interest Probability)
The probability of an interest x_{ki} can be defined as its frequency if divided by the cardinality of the community C_k:

$$P(x_{ki}) = \frac{if(x_{ki})}{|C_k|} \quad (2)$$

Definition 10: (Node interests' importance within a community)
The importance of the interests of a node n_j is a measure that specifies the importance of the user U_j in a determined community C_k knowing all of its interests. This measure is defined by the sum of $P(x_{ki})$ normalized by the size of the characteristic vector of U_j.

$$NI(n_j, C_k) = \frac{\sum_{i=1}^{D} P(x_{ki})}{D}. \quad (3)$$

Definition 11: (Overlapping node interests' importance)
Given an overlapping node v, $v \in \{C_1, ..., C_k\}$ where $k \geq 2$, importance of v interests is given by the average value of its NI in each community to which it belongs. This measure is defined by:

$$NI(v, C_{1,...,k}) = avg(NI(v, C_1), ..., NI(v, C_k))$$

These measures help us to state our diffusion model in the following but now in order to generate the candidate nodes, we will propose a similarity measure that assigns a weight to each edge of the graph in order to apply the link analysis algorithm PageRank.

Definition 12: (Similarity between two users)
The similarity between two users U_i and U_j specifies shared interests between them. It is given by:

$$Sim(U_i, U_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}. \quad (4)$$

where X_i and X_j are respectively the characteristic vectors of U_i and U_j.

2) Candidate nodes generation: After that, we seek candidate nodes from each community. For this purpose, we will apply the PageRank algorithm that takes as input a sub-graph (community) and gives us a measure of the importance of each node within this community. Based on this measure, we can estimate the candidate elements CN_i generated from the community i. So, the set of all candidate nodes of the graph G is $CN = CN_1 \cup CN_2 \cup, ..., \cup CN_K$ where K is the number of communities detected during the partition phase. Overlapping nodes generated by DOCNet are taken as candidates and are added by default to CN because of their important role in transferring information.

3) Influential nodes selection: Finally, once we have generated the set CN of candidates, we can select the set I of influential nodes. At this stage we apply a diffusion model that allows us to identify the likely nodes to disseminate information received (called active).

Proposed diffusion model:
We try to introduce the semantic aspect in our diffusion model, and inspired by the LTM [1], we can adopt the following model:
For each node v in a community C is associated a threshold value; this value is the weight $NI(v, C)$. Given the threshold values and an initial set of active nodes that corresponds to the set of candidates previously generated, a node v is influenced if the total weight of similarity with its active neighbors is at least $NI(v, C)$.
More formally:

$$\sum_{w \in N_v} sim(v, w) \geq NI(v, C). \quad (5)$$

where N_v is the set of active neighbors of v.
This model works iteratively and is repeated until no inactive node can be activated. At the end of the diffusion process we get a set of active nodes.

The previous steps have allowed us to study a local way network members and their relationships, to treat this problem globally we attribute for each node a label indicating the community to which it belongs. Then, we start searching influential nodes from the community C with the maximum number of actives nodes denoted by A_C. After that, we select the nearest node to all active members of its community (with the smallest value of influence distance), determine its area of influence and finally, delete this area from its community and pass to the next node until the community becomes empty.

Definition 13: (Shortest path between two nodes)
The shortest path between two nodes u and v in a graph G is a sequence of edges beginning in u and ending in v such that the sum of weight of these edges is minimized.
we will use *Dijkstra* algorithm to determine shortest paths.

Definition 14: (Distance between two nodes)
The distance between two nodes u and v of G denoted $dist(u, v)$ is the number of edges located in the shortest path that lead from u to v.

Definition 15: (Influence distance)
The distance of influence of a node u in a community C corresponds to the average distance of u to different active members of C, it is defined by the sum of distances from u to $v \in C$ divided by the cardinality of C.

$$dist_{inf}(u, C) = \frac{\sum_{v \in C} dist(u, v)}{|C| - 1} \quad (6)$$

The basic idea of the selection step is, to determine for each candidate node the shortest path leading to different active members of its community, then we can determine the closest one of the other nodes and thus has the more likely to maximize the influence in a well defined area called influence area, we assigned to this measure of centrality the name of *Farness Centrality*. Therefore, we can select the influential nodes that activate the maximum number of individuals in the network.

B. Properties

Property 1: For any node u and v from the graph G such that $u \neq v$ we have:

$$0 \leq sim(u, v) \leq 1$$

.

Property 2: Let v be a node belonging to a community C:

$$0 < NI(u, C) \leq 1$$

.

Property 3: Let u a node from the graph G and let C_i and C_j two communities such that $u \in C_i$ and $u \notin C_j$, so we have:

$$NI(u, C_j) = 0$$

.

Theorem 1: Let C be a community and let u be a node from A_C having the smallest influence distance, i.e:

$$arg(u) = min\{dist_{inf}(n_j, C)/n_j \in A_C\}$$

Let $\sigma(u)$ be the influence area of u, then any node that is part of $\sigma(u)$ can't have a smaller influence distance than u.

Demonstration 1: **If** $u \in C$ and $arg(u) = min\{dist_{inf}(n_j, C)/n_j \in A_C\}$, **then** $\forall v \in C$ such that $v \neq u$ we have $dist_{inf}(v, C) \geq dist_{inf}(u, C)$.
Now suppose that $v \in \sigma(u)$, therefore:

$$dist_{inf}(u, C) \leq dist_{inf}(v, C)$$

C. Algorithm

For this section, we give the sequence of steps as described above with the algorithm DIN.

DIN is outlined in Algorithm 1 this is explained in details in the following paragraph. We begin by detecting overlapping communities. Then, for each community we calculate the importance of interests of its nodes and the similarity between them. Once, we attribute a weight for each edge (line 10), we can apply PageRank algorithm and generate candidate nodes. A node is said candidate if its PageRank measure (PR) exceeds the average PageRank value of the members of its community (we note α in algorithm 2). Next step is the selection of influential nodes from the candidate ones. For each candidate, we determine the shortest path that leads to active members of his community (the farness centrality). After that, we order nodes by their distance of influence in an ascending order and start by the first node, determine his influence area, attribute it to I, then remove this area from the community and select the next node in Inf until the community becomes empty or there are no active nodes. A major difference of our model compared to others is that it returns the number of influential

Algorithm 1: Detecting Influential Nodes in social networks : DIN

Input: A graph $G = (V, E)$, a set of users' interests F
Output: A set of influential nodes I
begin
1 $I \leftarrow \emptyset$,$CN \leftarrow \emptyset$
2 Apply DOCNet algorithm to detect overlpping communities.
3 $O \leftarrow$ the set of overlapping nodes generated by DOCNet.
4 $CN \leftarrow CN \cup \{O\}$.
foreach *Detected community* C **do**
5 **foreach** *node* $u \in C$ **do**
6 Estimating a characteristic vector X_i of interests for each user.
7 Assign the importance of interests by applying the equation 3.
end
8 **foreach** *pair of nodes* (u, v) *of the community* C **do**
9 calculate the similarity $sim(u, v)$ using equation 4 .
10 $w(u, v) \leftarrow sim(u, v)$.
end
11 **Candidates generation** (C,G).
end
12 Apply the proposed diffusion model to obtain the set of active nodes.
13 **Influential nodes selection**(CN,G).
14 Return (I).
end

Algorithm 2: Candidates generation (C, G)

Input: A graph $G = (V, E)$, a community C,a threshold parameter α
begin
1 Apply PageRank algorithm to measure the importance of each node in C.
2 **foreach** *node* u *from the community* C **do**
3 **if** $(PR(u) \geq \alpha)$ **then**
4 $CN \leftarrow CN \cup \{u\}$
end
end
5 Return (CN).
end

Algorithm 3: Influential nodes selection (CN, G)

begin

Label each candidate node by his community.
$j \leftarrow 0$
while *(there are communities)* **do**
  choose the community C_K admitting the maximum number of active nodes.
  $A_{C_k} \leftarrow \{$ active nodes of $C_k\}$.
  while *($(C_k \neq \emptyset)$ and $(A_{C_k} \neq \emptyset)$)* **do**
    Calculate the influence distance of each element of $CN \in C_k$.
    $Inf \leftarrow$ Sort nodes in CN by their distance in ascending order.
    $v_{max} \leftarrow$ select the first node of Inf.
    Determine $\sigma(v_{max})$.
    $C_k \leftarrow C_k \setminus \{\sigma(v_{max})\}$.
    $I \leftarrow I \cup \{v_{max}\}$
    $j++$.
  end
  seed size $\leftarrow j$.
end
Return (I).

end

nodes (seed size) unlike other models where the seed size is an input to the model.

Theorem 2: The time complexity of our algorithm DIN is $O(n^2)$ where n is the number of nodes.

Proof. To calculate the time complexity of our algorithm, we will start by calculating the complexity of each phase. For the first phase of partition, we used DOCNet algorithm which admits a temporal complexity $O(n^2)$ [11].
To determine the time complexity of the second phase, we must determine the complexity of each step. First, we have assigned a weight to each node and each edge of our network which admits complexity of $O(m+n)$ where n is the number of nodes, and m is the number of edges.
Then, to the step of generating the candidates we have used the algorithm PageRank which admits a complexity of $O(n^2)$ and for each node we will compare his PR measurement with a certain threshold value which admits a complexity $O(n)$. The generation of candidates is repeated k times as k is the number of communities detected during the first phase, therefore this step admits a complexity of $O(kn^2)$.
Now, we should have the complexity of the third step of influential nodes selection. For this step, we start applying our diffusion model on our graph which admits a complexity of $O(n)$ in the worst case. Let c be the size of candidate nodes set, the next task is to calculate the shortest path of the candidates of a community in $O(m + clog(c))$. After performing this calculation, we will sort the result and put in a table, which admits a complexity of $O(clog(c))$ in the worst case and finally, we will choose the first item and determine its influence to happen next to all items stored in table Inf. This task then admits in the worst case a complexity of $O(c)$. Therefore, the time complexity of this selection step is:
$T_{selection} = O(max\{n, (clog(c)), m + clog(c)\} = O(m + clog(c))$
we conclude that the time complexity of the second phase of our algorithm is :

$$T_{phase2} = O(max\{m+n, kn^2, m+clog(c)\}) = O(kn^2)$$

Note finally that in the general case, the number k of communities is very negligible compared to n so the time complexity of DIN is $O(n^2)$.

Theorem 3: The space complexity of our algorithm DIN is $O(nd_{max})$ where n is the number of nodes and d_{max} is the dimension of the largest characteristic vector.

Proof. In our algorithm, we estimate a characteristic vector of interests for each network user. This vector is associated to each vertex of graph which generates a spatial complexity of $O(nd_{max})$ where n is the number of vertices of the graph and d_{max} is the dimension of the largest characteristic vector.

5. Experimental results

The main purpose of this section is to compare experimentally our proposal with two well-known algorithms: 1) **FBI** [13] part of approaches based on the Pagerank algorithm and 2) **Coreness Centrality** [18] which introduces a new centrality measurement to estimate the influence of a node based on the centrality index of its neighbors. But before reporting the results, we need to indicate the performance indices that we will adopt and the used networks.

A. Evaluation metrics

The evaluation of the performance of maximizing influence in a network is not trivial. We consider two metrics:

1) Influence spread: Influence spread is a metric to measure the number of nodes that can be influenced by the set of k influential nodes. Our proposal is parameterless, it indicates the number of influential nodes and those influenced as output. For the other models, we make the execution several times and choose $top - k$ influential nodes that provide maximum influence spread.

2) Running time: Most existing studies that try to detect influential elements suffer from the time they supply to run their models particularly those based on a greedy algorithm.

We considered for evaluation artificial networks whose structure (communities) is well known.

B. Artificial networks

To evaluate our proposal and the two others models, we adopted the LFR benchmark [19] for synthetic networks. In our experiments, we used networks composed of N=1000; N=2000, N=5000, N=8000, N=10000 and N=20000 nodes [1].

The rest of the parameters are as follows: The average degree is kept at $k = 10$; the mixing parameter μ is 0.2; the

[1]The maximal number of nodes was limited by the capacity of our machine consisting of (Intel(R)Core(TM)i3- 3227U CPU 1.90 GHz with 4 Go of memory).

maximum degree is 50; the community size varies between 20 and 100; O_n (the number of overlapping nodes) is set to 10%, O_m (the number of communities to which each overlapping node belongs) is set to 2. Before setting these parameters we seek to show the results obtained with the variation of network complexity.

1) DIN performance with the variation of network complexity: we generated a set of graphs by varying the average degree of a node as well as varying the mixing parameter μ which denotes the fraction of edges outside the community of each node relative to the total number of edges. The benchmark contains 1000 nodes, the community size ranges from 20 to 100 and the mixing parameter μ is 0.2. O_m parameter is set to 2 and O_n is set to 10% (100 overlapping nodes).

Figure 2 illustrates the results while varying the average degree. In fact, we can see a significant result with Coreness centrality because such model is based on the centrality of a node's neighbors to determine its centrality, then the influence spread of a node increases by the number of its neighbors i.e degree, whereas our model keeps close values. Looking at the FBI curve, we notice that at a degree $k = 10$ the influence spread starts decreasing. We can conclude that DIN generates a good number of influenced nodes regardless of increasing the degree.

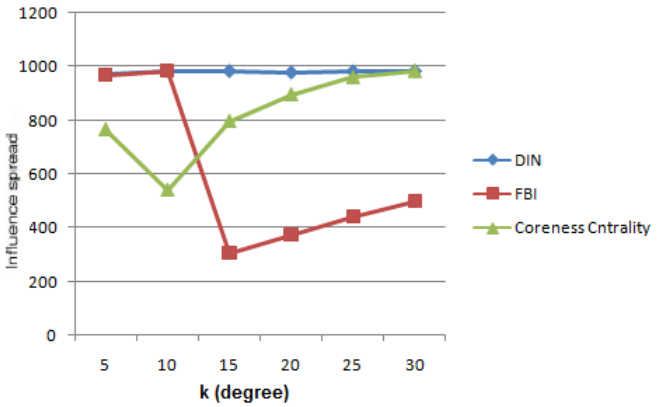

Fig. 2. Influence spread variation in function of k

For the running time, simulation results are reported in figure 3. According to the curves, we notice that FBI represents a limited runnig time thus this model conserves the best values while varying the degree. Our model maintains close values to FBI while the Coreness centrality representes high values.

As a second test for artificial networks while varying network complexity, we generated graphs with diferent values of the mixing parameter μ. Experimental results are shown in figure 4. We observe that FBI has the best influence spread despite that it is not based on a community structure at the detection of influential nodes but it is based on link analysis because of the use of the PageRank algorithm. The second approach is DIN, it gives close values to those of FBI. Coreness centrality has the lowest values. We can conclude that DIN offers significant results when the community structure

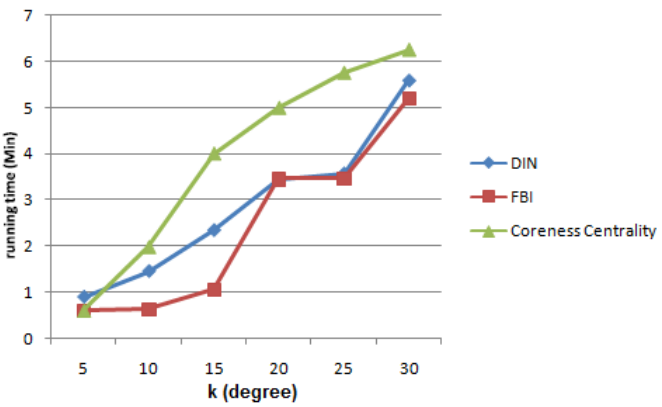

Fig. 3. Running time variation in function of k

is clear (i.e when μ increases). Similar results are obtained for the runnig time while varying μ. As shown in figure 5 FBI has the lowest values followed by DIN then Coreness Centrality.

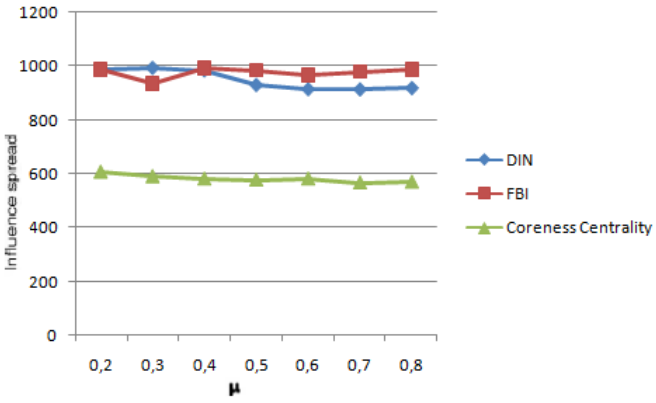

Fig. 4. Influence spread variation in function of μ

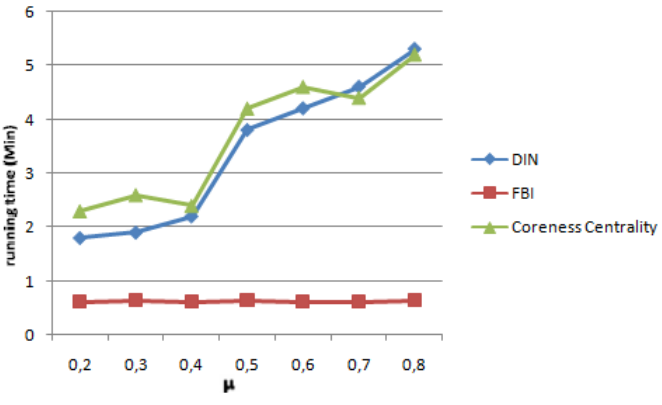

Fig. 5. Running time variation in function of μ

2) DIN performance with the variation of network size: In this second part of simulations, we consider LFR graphs

containing 1000, 2000, 5000, 8000, 10000 and 20000 nodes. Experimental results are summarized in Tables I and II.

TABLE I
THE VALUES OF THE INFLUENCE SPREAD WITH THE VARIATION OF THE NETWORK SIZE

N	**FBI**	**Coreness centrality**	**DIN**
1000	938	574	983
2000	969	1145	1962
5000	4675	2850	4877
8000	7790	4747	7827
10000	9834	5805	9784
20000	18585	11625	19557

TABLE II
THE VALUES OF THE RUNNING TIME(MINUTES) WITH THE VARIATION OF THE NETWORK SIZE

N	**FBI**	**Coreness centrality**	**DIN**
1000	1.02	2	1
2000	1.13	9	9
5000	13	32	35
8000	22.2	294	70
10000	32.5	415	229
20000	140.5	1645	974

For the LFR graphs while varying the size and from Table I we find that with the change of the network size, our model is always better than FBI and Coreness Centrality for the influence spread. With a few number of influential elements, DIN covers a large number of influenced nodes in the network. As shown in the results, our proposal is able to have better value of influence spread while increasing the network size. On the other hand, for the running time and from Table II we notice that FBI is the best one followed by DIN then the Coreness centrality model. For our model, the partition phase takes time that's why with large graphs it reaches high values.

6. CONCLUSION

In this paper, we focus on the problem of influence maximization in social networks. Many current researches on this problem are developed and we have discussed their limits. Our main contribution is the proposition of an algorithm called DIN (Detecting Influential Nodes in social networks) which combines the structure and the semantic of the network. We have compared our proposal with two well known approches and we have used artificial networks as datasets. Simulation results are very encouraging and show the great performance of our model in computing an optimal set of influential nodes having the maximal influence spread. For the time being, our immediate concern is to test our model on real-world large scale social networks from facebook. Unfortunately, such data is not publically available due to privacy. We are actually contacting several social networks to get such data which we will anonymize to preserve personal information.

REFERENCES

[1] K. Rahimnkhani, A. Aleahmed, M. Rahgozar and M. Moeini. *A fast algorithm for finding most influential people based on the linear threshold model*, Expert Systems with Applications, vol. 42, 2015, pp. 1353-1361.

[2] Cedric Lagnier and Eric Gaussier , *Etude de la Maximisation de lInfluence dans les Rseaux Sociaux*, MARAMI, 2013.

[3] D. Kempe, J. Kleinberg and E. Tardos, *Maximizing the spread of influence through a social network*, In Paper presented at the ninth ACM SIGKDD international conference on knowledge discovery and data mining. Washington, D.C., 2003, pp. 137-146

[4] Z. Tian , W. Bai , W. Bin and Z. Chuanxi, *Maximizing the spread of influence ranking in social networks*, Information Sciences, vol. 278,2014, pp. 535-544.

[5] P. Domingos and M. Richardson, *Mining the network value of customers*, in: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, pp. 57-66.

[6] Y. Chen, W. Zhu, W. Peng, W. Lee, and S. Lee. *CIM : Community-based influence maximization in social networks*, ACM Transactions on Intelligent Systems and Technology (TIST) Special Issue on Linking Social Granularity and Functions, 5(2), 2014.

[7] M. Heidari, M. Asadpour, and H. Faili, *SMG :Fast scalable greedy algorithm for influence maximization in social networks*, Physica A, vol. 420, 2015, pp. 124-133.

[8] A. Goyal, W. Lu, and Laks V.S. Lakshmanan.*CELF++ : Optimizing the greedy algorithm for influence maximization in social networks*, 2011.

[9] R. Walia, G. Sarna, and M.P.S. Bhatia. *Finding the influential overlap nodes in communities*, In Computational Intelligence Communication Technology(CICT), 2015 IEEE International Conference on, , 2015, pp. 349-353.

[10] A. Lancichinetti, S. Fortunato and J. KertTsz, *Detecting the overlapping and hierarchical community structure in complex networks*, New Journal of Physics, 11(3), 2009.

[11] Delel Rhouma and Lotfi Ben Romdhane, *An efficient algorithm for community mining with overlap in social networks*, Expert Systems with applications, vol. 41, 2014, pp. 4309-4321.

[12] L. Grgano, H. Hell, J.G. Peters, and U. Vaccaro, *Influence diffusion in social networks under time window constraints*, Theoretical Computer Science, vol. 584, 2015, 53-66.

[13] G. Wang, W. Jiang, J. Wu, and Z. Xiong, *Fine-grained feature-based social influence evaluation in online social networks*, Parallel and Distributed Systems, IEEE Transactions on, vol. 25(9), 2014, pp. 2286-2296.

[14] F. Quan, S. Jitao, X. Changsheng and R. Yong, *Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning*, Multimedia, IEEE Transactions on, vol. 16(3), 2014, pp. 796-812.

[15] W. Xiaohui , J. Jia, T. Jie , W. Boya , C. Lianhong and X. Lexing, *Modeling emotion influence in image social networks*, Affective Computing, IEEE Transactions on, vol. 6(3), 2015, pp. 286-297.

[16] D. Myungcheol and L. Ling, *Probabilistic diffusion of social influence with incentives*, Services Computing, IEEE Transactions on, vol. 7(3), 2014, pp. 387-400.

[17] Sergey Brin and Lawrence Page, *The anatomy of a large-scale hypertextual web search engine*, Computer Networks and ISDN Systems, vol. 30(1-7), 1998, pp. 107-117.

[18] B. Joonhyun and K. Sangwook, *Identifying and ranking influential spreaders in complex networks by neighborhood coreness*, Physica A, vol. 395, 2014, 549-559.

[19] Benchmark graphs to test community detection algorithms, http://sites.google.com/site/santofortunato/inthepress2.