<table>
<thead>
<tr>
<th>Title</th>
<th>Combining intensification and diversification to maximize the propagation of social influence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Fan, X; Li, VOK</td>
</tr>
<tr>
<td>Citation</td>
<td>The 2013 IEEE International Conference on Communications (ICC 2013), Budapest, Hungary, 9-13 June 2013. In IEEE International Conference on Communications, 2013, p. 2995-2999</td>
</tr>
<tr>
<td>Issued Date</td>
<td>2013</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10722/191607">http://hdl.handle.net/10722/191607</a></td>
</tr>
<tr>
<td>Rights</td>
<td>IEEE International Conference on Communications. Copyright © IEEE.; ©2013 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.; This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
Combining intensification and diversification to maximize the propagation of social influence

Xiaoguang Fan*, Victor O.K. Li†
Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
*xfan@eee.hku.hk †vli@eee.hku.hk

Abstract—In this paper we consider the influence maximization problem in social networks, and propose an Int-Div heuristic to solve it. Motivated by the concepts of intensification and diversification in optimization problems, Int-Div accounts for both of these two concepts to estimate the social influence, and selects nodes based on marginal influence increment. It is applicable to the two widely used diffusion models, namely, the Linear Threshold Model and the Independent Cascade Model. The proposed strategy is evaluated through experiments on a collaboration network and a who-trust-whom online social network, respectively, and compared with several existing heuristics, namely, the pure greedy algorithm, the centrality-based scheme, the single discount and the degree discount heuristics. We find that our proposed strategy offers better performance than the centrality-based scheme, the single discount and the degree discount heuristics, while achieving approximately the same performance as the greedy algorithm. The computational load is dramatically lower than the greedy heuristic.

I. INTRODUCTION

Information propagation in social network has attracted much attention from IT researchers, sociologists and economists. It has been analyzed in various settings. [1] studies the evolution of node status by the Markov Chain model. [2] uses game theory to analyze innovation spreads. [3] utilizes person-to-person recommendation networks to study the information propagation and further presents an efficient viral marketing model. [4] evaluates the adoption of information technologies in the context of social computing. Moreover, [5] utilizes the local mean field technique to model the influence spreading process, and obtains the recursive distributional equations for the diffusion model. Comprehending how to maximize the influence propagation is critical in advertising, marketing, political movements, and new technology adoption. With the rapid development of online social networks and social medias like Facebook, Twitter and LinkedIn, large-scale, high-speed and instantaneous information dissemination becomes possible, and many business opportunities are available, such as viral marketing and social coupon. For example, [6] presents a practical marketing strategy specifically designed for HokeyPokey, a super-premium ice-cream retailer. Social media marketing may potentially spawn enormous business opportunities for electronic commerce.

We define the influence maximization problem as below: If \( K \) nodes are permitted to be activated initially, how do we select them so that the maximum spreading of activation is obtained in terms of the final active set size? [7] [8] first illustrate the issue as an algorithmic problem. It is further formulated as an NP-hard optimization problem and a greedy heuristic is proposed with a provable approximation guarantees of the optimal in [9]. The greedy algorithm calculates the influence power precisely by enumeration. The more rounds the simulation takes, the more accurate the result will get. However, when the network size increases, the computational time will increase dramatically, rendering the greedy algorithm infeasible for the influence maximization problem in the real world. [10] discounts the degree of each node by removing the neighbors that are already in the active seed set, and proposes the single discount heuristic. In addition, for the Independent Cascade Model [11] with small propagation probability, the discount on degree value could be calculated in detail. This leads to the upgraded strategy called degree discount heuristic. However, the underlying methodology of the degree discount heuristic limits its applicability only to the Independent Cascade Model with a very small propagation probability.

Intensification and diversification are key concepts in metaheuristics for combinatorial optimization problems [12]. These two concepts and their underlying strategies greatly determine the structure of a metaheuristic. In order to maximize the social influence, we should select the nodes that could cooperatively trigger a big cascade of activations while reducing overlapping activations among them as much as possible. We borrow the concepts of intensification and diversification in optimization problems, and apply them to our problem, by designing an Int-Div heuristic which strikes a balance between them. Our contributions include (a) proposing an Int-Div algorithm that combines both intensification and diversification to estimate the influence spread; (b) demonstrating that our strategy achieves approximately the same performance as the greedy algorithm, but with much lower computational cost, and that it also obtains better performance than the degree-based scheme, the single discount and degree discount heuristics, with the cost staying in the same order of magnitude with them; (c) showing the scalability of our scheme across various datasets with different diffusion models.

II. PROBLEM FORMULATION

A. Problem description

Considering the target consumer group as a directed social graph \( G(V,E) \) where \( V \) is the set of vertices and \( E \) is the set of directed edges, we define the in-degree and out-degree neighbor set of Node \( j \) as \( N_{in}^j = \{ i \in V : (i,j) \in E \} \) and \( N_{out}^j = \{ i \in V : (j,i) \in E \} \), respectively. Here edge \( (i,j) \) is
directed from Node \(i\) to Node \(j\). Each node is active or inactive, and the probability for a node switching from inactive to active increases as more of its in-degree neighbors become active. Note that an active node cannot return to the inactive state. Originally all of the nodes are inactive, and we launch our marketing strategy to make \(K\) nodes active. In the following steps, nodes are activated by their active in-degree neighbors and in turn activate their inactive out-degree neighbors. The process is finished when no more activation is possible. The influence maximization problem is as follows: If \(K\) nodes are permitted to be activated initially, find this \(K\)-node set to maximize the final expected number of active nodes.

B. Diffusion models

Two basic diffusion models, namely, the Linear Threshold Model (LTM) [13] and the Independent Cascade Model (ICM) [11], are presented to determine how a node is influenced by its neighbors. Both models evolve in discrete time steps. In LTM, a Node \(j\) has a weight \(b_{i,j}\) on edge \((i,j)\) to express the influential power on Node \(j\) and \(\sum_{i\in N_{in}^{j}} b_{i,j} \leq 1\). A threshold \(\theta_j\) is pre-assigned to Node \(j\), and it is uniformly distributed in \([0,1]\). At any single step, if the weight summation of all the active in-degree neighbors of Node \(j\) exceeds \(\theta_j\), Node \(j\) is activated. In ICM, Node \(i\) that becomes active at Step \(t\) has a probability \(p_{i,j}\) to successfully activate each inactive out-degree neighbor \(j\) through edge \((i,j)\) at Step \(t+1\). The probability is independent of the historical activation track, and \(i\) does not have any chance to activate \(j\) again whether or not it succeeds. Note that the order of activation is random when multiple active neighbors try to activate a common node.

III. METHODOLOGY

Suppose the influence power of each node could be divided into two independent parts, namely, the intensification and diversification parts. From the two diffusion models we mentioned in II-B, the probability that a node is activated increases when the number of its active neighbors goes up. It means that these active nodes work as a team to achieve intensification on activating their common neighbors. However, if we focus excessively on the teamwork effect, the active seeds may be clustered, and they may form a group in which seeds are neighbors of each other. In this case, the influential range of each cluster is pre-assigned to Node \(j\) at Step \(t\) has a weight \(\theta_j\) on edge \((i,j)\) to express the influential power on Node \(j\) and \(\sum_{i\in N_{in}^{j}} b_{i,j} \leq 1\). A threshold \(\theta_j\) is pre-assigned to Node \(j\), and it is uniformly distributed in \([0,1]\). At any single step, if the weight summation of all the active in-degree neighbors of Node \(j\) exceeds \(\theta_j\), Node \(j\) is activated. In ICM, Node \(i\) that becomes active at Step \(t\) has a probability \(p_{i,j}\) to successfully activate each inactive out-degree neighbor \(j\) through edge \((i,j)\) at Step \(t+1\). The probability is independent of the historical activation track, and \(i\) does not have any chance to activate \(j\) again whether or not it succeeds. Note that the order of activation is random when multiple active neighbors try to activate a common node.

A. Marginal influence increment

In order to predict the influence spread, we define a metric called the marginal influence increment, which evaluates the marginal gain of activation one node could make based on a given active seed set. Suppose \(S\) is the existing active node set, we hope to select the next target node \(i\) and add it to \(S\). Intuitively we select the node that makes the maximum marginal increase of influence. If \(p_i(S)\) represents the probability that Node \(i\) is influenced by node set \(S\), then with probability \((1 - p_i(S))\) \(S\) fails to activate Node \(i\). In this situation, the marginal influence gain, or the additional nodes activated by selecting Node \(i\) into \(S\) consists of the following three parts: 1) Node \(i\) itself; 2) the intensification part \(\text{INT}_i(S)\), which corresponds to the common out-degree neighbors influenced cooperatively by Node \(i\) and seed set \(S\); 3) the diversification part \(\text{DIV}_i(S)\), which corresponds to the nodes who are the out-degree neighbors of Node \(i\), but not the out-degree neighbors of any node in \(S\). Thus the marginal influence increment \(\text{MII}_i(S)\) of Node \(i\) based on the active seed set \(S\) could be calculated as follows:

\[
\text{MII}_i(S) = (1 - p_i(S))(1 + \text{INT}_i(S) + \text{DIV}_i(S))
\]

There are two different sets of equations to calculate \(p_i(S)\), \(\text{INT}_i(S)\), and \(\text{DIV}_i(S)\) in LTM and ICM, respectively.

For LTM:

\[
p_i(S) = \sum_{j\in S, j\in N_{in}^i} b_{i,j} \\
\text{INT}_i(S) = \sum_{j\notin S, j\in N_{out}^i, \exists r\in S} b_{i,j} \cdot (1 + \sum_{r\in S} b_{j,r}) \\
\text{DIV}_i(S) = \sum_{j\notin S, j\in N_{out}^i, \forall r\in S} b_{i,j} \cdot (1 + \sum_{r\notin S} b_{j,r})
\]

For ICM:

\[
p_i(S) = 1 - (1 - p)^k \\
\text{INT}_i(S) = \sum_{j\notin S, j\in N_{out}^i, \exists r\in S} [(1 - p)(N_{in}^j \cap S) - (1 - p)(|N_{in}^j \cap S| + 1)]; \\
\text{DIV}_i(S) = \sum_{j\notin S, j\in N_{out}^i, \forall r\in S} p \cdot [1 + p \cdot (d_{j\in N_{out}^i} - 1)]
\]

where \(k\) is the number of active in-degree neighbors of Node \(i\), \(d_{j\in N_{out}^i}\) is the out degree of Node \(j\), and \(|N_{in}^j \cap S|\) is the number of elements in the set \(N_{in}^j \cap S\). B. up-to-2-hop degree metric

While evaluating the centrality of node, degree may be the obvious choice since it directly demonstrates one’s connectivity. However, we should also consider the neighbors’ influence when evaluating one’s influential power. Considering the notion of weak tie [14] [15], which emphasizes the relationship between people with common friends, we assume that 2-hop neighbors are still susceptible nodes in the influence spreading process. We define an up-to-2-hop degree metric \(\mu_i\) to evaluate the individual influence of Node \(i\). \(\mu_i\) is the sum of weighted degrees of each node whose social distance is at

\footnote{The social distance is defined as the number of edges on the shortest path between two nodes.}
most 2 from Node \( i \). In LTM, \( \mu_i \) is calculated as follows:

\[
\mu_i^{\text{LTM}} = \sum_{j \in N_{r_{i}}^\text{o}} b_{i,j} + \left( \sum_{j \in N_{r_{i}}^\text{o}} b_{i,j} \right) \cdot \left( \sum_{r \in N_{r_{i}}^\text{o}, r \neq i} b_{i,r} \right) \tag{4}
\]

where \( N_{r_{i}}^\text{o} \) is the out-degree neighbor set of Node \( i \), and \( b_{i,j} \) is the weight that evaluates how much Node \( i \) could influence its neighbor \( j \). Notice that in the right side of the equation, the first addend is the 1-hop influence of Node \( i \), while the second addend represents the 2-hop influence. Similarly, we have an expression of \( \mu_i \) in ICM:

\[
\mu_i^{\text{ICM}} = p \cdot d_{i}^{\text{o}} + \sum_{j \in N_{r_{i}}^\text{o}} p^2 \cdot (d_{j}^{\text{o}} - 1) \tag{5}
\]

where \( d_{i}^{\text{o}} \) is the out-degree of Node \( i \), and \( p \) is the activation probability.

C. Int-Div heuristic

The whole process of our Int-Div heuristic could be illustrated as follows: Given a sequence of nodes sorted by decreasing up-to-2-hop degree metric, we need to select \( K \) nodes. The head of the sorted list is selected at first and added into the existing active node set \( S \). Next we calculate the marginal influence \( MII(S) \) for each node not in \( S \), and add the node \( i \) with the maximum \( MII(S) \) into \( S \). In the following steps, we check the rest of the nodes in the same way. The process ends when \( K \) nodes are found.

In order to better illustrate how the proposed algorithm works and how \( MII(S) \) is calculated, we use an example to illustrate the seed selection process. Figure 1 and Tables I show the social graph and calculation results of \( MII(S) \) in LTM, respectively. The values of selected nodes are marked with bold font. Suppose we want to select 4 nodes. Obviously node \( C \) is chosen at first (\( \mu_C = \left( 2 \cdot \frac{3}{6} + \frac{1}{2} \right) + \frac{1}{2} \cdot \frac{3}{6} = 3.833 \)). The next choice should be node \( A \), since \( MII_A(S) = 2.667 \) is maximal. When calculating \( MII_i(S) \), there are two parts \( INT_A(S) \) and \( DIV_A(S) \). Node \( A \) has one common neighbor \( B \) with the activated node \( C \), and one separate neighbor \( E \). The intensification part should consider the additional influence \( A \) would make on \( B \) under the activation of \( C \), while the diversification part is concerned with the independent activation ability from \( A \) to \( E \) out of the influence range of \( C \). According to Equation 2, \( INT_A(S) = \frac{1}{2} \cdot \left( 1 + 1 \right) = 0.667 \) and \( DIV_A(S) = 1 \). Thus \( MII_A(S) = (1 - 0) \cdot (1 + 0.667 + 1) = 2.667 \). Note that although Node \( B \) has the second largest up-to-2-hop degree metric (\( \mu_B = 2.875 \)), it is not selected since \( MII_B(S) = \left( 1 - \frac{1}{2} \right) \cdot \left( 1 + 1 \right) = 2 < MII_A(S) \). In the following steps, Nodes \( I \) and \( F \) are selected since \( MII_I(S) = 1.5 \) and \( MII_F(S) = 1.333 \). Here we find that the additional activation capabilities of Nodes \( B \) and \( D \) decrease dramatically after Nodes \( C \), \( A \), and \( I \) are selected, since the influence ability is reduced significantly by \( (1 - \mu_i(S)) \) for the candidate nodes who are neighbors of the active node set \( S \). Similarly, we can deduce the algorithm for ICM\(^2\).

\(^2\)The calculation is similar and we left it out due to space limitations.

Finally, a formal graph in the Linear Threshold Model.

**Algorithm 1: Int-Div heuristic**

**Input:**
- Let 1,...,\( N \) be nodes. We need to select \( K \) seeds.
- LTM: Network \( G(V,E) \) with weight \( b_{i,j} \) for \( (i,j) \subseteq E \).
- ICM: Network \( G(V,E) \) with activation probability \( p \).

**Output:**
- The active seed set \( S \).
  1. Start with \( S = \emptyset \).
  2. Calculate \( \mu_i \) for each node \( i (i = 1 \text{ to } N) \).
  3. Select \( u = \text{argmax}_i \{ \mu_i \} \).
  4. \( S \leftarrow S \cup \{ u \} \).
  5. while \( |S| < K \) do
     a. Calculate \( MII_i(S) \) for node \( i (i = 1 \text{ to } N \text{ and } i \notin S) \).
     b. Select \( v = \text{argmax}_i \{ MII_i(S)|i \notin S \} \).
     c. \( S \leftarrow S \cup \{ v \} \).
   end while

**IV. EXPERIMENTAL RESULTS**

In this section we evaluate our Int-Div heuristic with four typical strategies in LTM and ICM, respectively. We
implement a real collaboration network and a who-trust-whom online social network to test the algorithms.

A. Dataset

- The collaboration network (CN) is a co-authorship network of Arxiv General Relativity [16]. Here we assume that both the edges \((i,j)\) and \((j,i)\) are included in the graph if author \(i\) and \(j\) co-authored one paper, and multiple paper co-authorships between the same pair of authors will not cause more edges. Note that this co-authorship network is a symmetric graph. The largest connected component is utilized to build a graph with 4158 nodes and 26850 edges.

- The Epinions social network (ESN) is a who-trust-whom online social network from a review site Epinions.com [17]. Users of the website can determine whether to “trust” others. The trust relationships will form the Web of Trust which helps to rank the reviews for the users. If User \(i\) chooses to trust User \(j\), a directed edge from Node \(j\) to Node \(i\) is included in the graph. Note that this trust network is an asymmetric graph. In other words, an edge from Node \(i\) to Node \(j\) may not guarantee the existence of an edge from \(j\) to \(i\), since most of the time trust is not bi-directional behavior. We consider the largest connected component of this dataset. It has 75879 nodes and 508837 edges.

These two datasets provide various social network structures (symmetric vs asymmetric) with different network scales (4158 nodes vs 75879 nodes) from two distinct areas (academic research vs real life).

B. Simulation setting

- Int-Div heuristic is proposed by us. It could be used in all the cases and for both LTM and ICM.

- Pure greedy strategy [9] uses hill-climbing heuristic to evaluate the influence of each node by enumeration.

- Degree-based scheme such as [9] selects nodes in the order of decreasing out-degree values.

- Single discount heuristic [10] makes the discount on the degree of each node by removing the neighbors that are already in the active seed set, then selects the node with the maximal discounted degree value at each step. It is applicable in both LTM and ICM.

- Degree discount heuristic [10] is proposed for ICM with small propagation probability \(p\). Based on the single discount heuristic, it discounts the degree value in detail, since the indirect influence of a node to multi-hop neighbors could be ignored by the assumption of small \(p\). As [10] shows that it still has a good performance in LTM model, we will report results for both of the two diffusion models except ICM with large \(p\) such as \(p = 20\%\). Moreover, we assign \(p = 0.5\%\) to calculate the degree discount value in LTM.

The performance metric is the number of active nodes at the end of the process. We take 1000 runs for each initially active seed set and calculate the average.

C. Result

Figures 2 and 3 show the performance comparison of the Int-Div heuristic (IntDiv), the greedy algorithm (Greedy), the degree-based scheme (Degree), the single discount (Single Discount) and the degree discount strategies (Degree Discount) in the LTM for CN and ESN datasets, respectively. The x-axis represents the number of initial active seed set size. In CN dataset, our Int-Div heuristic gives the best performance of the five algorithms compared except when the initial seed set size is less than 15, when the greedy heuristic is better. In ESN dataset, we only record the results for initial seed set size up to 12 for the greedy heuristic, since the computation time is already almost 58 days and we could assume that it is computationally infeasible when the size is bigger than 12. Even focusing on seed set size up to 12, the performance of our heuristic is approximately the same with the greedy algorithm. Moreover, it is the best among the other four schemes.

Figures 4 and 5 evaluate the performance of the five algorithms in ICM with probability \(p = 5\%\) and \(p = 20\%\) for CN dataset, respectively\(^3\). Here we do not report the results from the degree discount heuristic when \(p = 20\%\) since

\(^3\)ESN dataset has similar results and we left it out due to space limitations.
it is not applicable to the situation with large propagation probability. We find that our heuristic performs the best except when the initial seed set size is less than 36 with \( p = 5\% \), when the greedy heuristic is better. For the case \( p = 20\% \), our heuristic is better than the single discount heuristic and the degree-based scheme. One observation is that when the activation probability is big enough (20\%), the first selected node makes the majority of activations in all of the algorithms. The subsequently selected nodes make nearly no progress for the centrality-based schemes due to the problem of overlapping influence ranges of clustered nodes mentioned in Section III. However, our Int-Div heuristic keeps enhancing the performance since it utilizes the diversification mechanism \( DIV(S) \) to avoid overlapping. Moreover, the difference on performance with the greedy algorithm is less than 6\%. When the initial seed size is 100, the difference is only 1.8\% of the greedy algorithm. Considering the huge computational savings (see Table II), this small difference on performance is acceptable.

Next we discuss the program running time. Table II shows the average computational time for each algorithm, using a Quad-core 3.0GHz PC with 32-bit operating system. The running time for the Int-Div heuristic is just 0.125% and 0.078% of the greedy algorithm in LTM for CN and ESN datasets, respectively. Meanwhile the computation time of our heuristic is only 0.179%, 0.142% of the greedy algorithm in ICM for CN dataset with \( p = 5\% \) and \( p = 20\% \), respectively. Moreover, the cost of our heuristic is within the same order of magnitude with the degree-based scheme, the single discount and the degree discount heuristics. In fact, for large datasets such as ESN, the greedy algorithm is computationally infeasible. Therefore, our proposed heuristic is probably the best to obtain high performance with limited cost.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Greedy</th>
<th>Int-Div</th>
<th>Degree</th>
<th>DegDis</th>
<th>SingDis</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTM(CN)</td>
<td>15.176</td>
<td>0.0228</td>
<td>0.02</td>
<td>0.0203</td>
<td>0.0201</td>
</tr>
<tr>
<td>LTM(ESN)</td>
<td>1.375</td>
<td>0.076</td>
<td>0.876</td>
<td>0.879</td>
<td>0.875</td>
</tr>
<tr>
<td>ICM(CN,5%)</td>
<td>5.129</td>
<td>0.0092</td>
<td>0.0051</td>
<td>0.0064</td>
<td>0.0052</td>
</tr>
<tr>
<td>ICM(CN,20%)</td>
<td>9.977</td>
<td>0.0142</td>
<td>0.0117</td>
<td>-</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Table II

COMPUTATION TIME FOR THE ALGORITHMS (HOURS)

V. FUTURE WORK

In this paper, the Int-Div heuristic is proposed to solve the influence maximization problem. In the future, we hope to build an analytical framework to quantitatively evaluate the performance of various influence maximization strategies.

REFERENCES


Figure 4. Performance for ICM with probability 5\% in CN dataset

Figure 5. Performance for ICM with probability 20\% in CN dataset